Essays on the Economics of Linguistic Diversity and Preference for Surprise

by

Haiyun Chen

M.A., McMaster University, 2009
B.A. (Hons.), University of Western Ontario, 2008

Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in the Department of Economics Faculty of Arts and Social Sciences

© Haiyun Chen 2017
SIMON FRASER UNIVERSITY
Summer 2017

All rights reserved. However, in accordance with the Copyright Act of Canada, this work may be reproduced without authorization under the conditions for “Fair Dealing.” Therefore, limited reproduction of this work for the purposes of private study, research, education, satire, parody, criticism, review and news reporting is likely to be in accordance with the law, particularly if cited appropriately.
## Approval

**Name:** Haiyun Chen  
**Degree:** Doctor of Philosophy (Economics)  
**Title:** *Essays on the Economics of Linguistic Diversity and Preference for Surprise*

**Supervisory Committee:**  
**Chair:** Simon Woodcock  
Chair  
Associate Professor, Department of Economics

**Dr. Arthur J. Robson**  
Senior Supervisor  
Professor, Department of Economics

**Dr. Gregory Dow**  
Supervisor  
Professor, Department of Economics

**Dr. Songzi Du**  
Supervisor  
Assistant Professor, Department of Economics

**Dr. Shih En Lu**  
Supervisor  
Assistant Professor, Department of Economics

**Dr. David Freeman**  
Internal Examiner  
Assistant Professor, Department of Economics

**Dr. B. Curtis Eaton**  
External Examiner  
Professor, Department of Economics, University of Calgary

**Date Approved:** July 19, 2017
The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016
Abstract

This thesis is composed of three essays, the first two of which are on the economics of linguistic diversity and the last on the evolutionary foundation of the preference for surprise.

Chapter 1 is joint work with Leanna Mitchell. We propose a theory that relates linguistic diversity to cooperative and competitive incentives in a game theoretic framework. In our model, autonomous groups interact periodically in games that represent either cooperation, competition, or no interaction. Language common to a pair of groups facilitates cooperation; whereas language unique to one group affords that group an advantage in competitions against other groups. The relative frequency of cooperation and conflict in a region provide incentives for each group to modify their own language, and therefore leads to changes in linguistic diversity over time. Our model predicts that higher frequency of cooperation relative to conflict reduces a region’s linguistic diversity.

Chapter 2 reports a laboratory experiment designed to test the theory proposed in the previous chapter. In the experiment, pairs of subjects endowed with a set of words interact repeatedly in a series of underlying games, in which they use the words to signal their intended action. The underlying games are either coordination or zero-sum. As the subjects are allowed to modify their vocabularies by learning words from their counterpart and creating new words, I observe that, over time, the pairs of vocabularies in coordination games tend to converge, while in zero-sum games, the vocabularies experience constant pressure to diverge. This finding is consistent with the theoretical predictions in Chapter 1.

Chapter 3 uses a principal-agent model to provide an evolutionary explanation of the preference for surprise, where surprise is measured by the Kullback-Leibler divergence. The principal in the model is interpreted as the blind force of evolution, who tries to maximize the fitness of the agent. The agent—generations of human beings—seeks to maximize a utility function designed by the principal. In a typical period, the agent first exerts a costly effort to gather information about an unknown state and then takes an action with respect to it. The variance of the signal distribution changes across time, but the agent is predisposed to believe that it does not. I show that if the variance of the signal distribution decreases at a sufficiently fast rate over time, it is evolutionarily optimal for the utility function to include a component that rewards surprises.
Keywords: economics; language; linguistic diversity; cooperation; conflict; preference for surprise; statistical decision theory; information acquisition
Dedication

To my parents, Chen Chengqi and Li Guoe, and my aunt Li Guoping.
Acknowledgements

I would like to thank, first and foremost, my senior supervisor Arthur Robson for his guidance, encouragement and support throughout my Ph.D. studies. Besides reading through countless drafts of my papers and providing valuable comments, Arthur also funded my experimental project which eventually becomes a chapter in this dissertation.

I am also grateful to the members of my supervisory committee, Greg Dow, Songzi Du, and Shih En Lu, who have been a rich source of enlightenment for me. Over the years, I have benefited enormously from the many insightful conversations with Greg, Songzi, and Shih En about research, teaching and career choices.

In addition, I extend my sincere gratitude to faculty members of SFU’s Economics Department, particularly, Doug Allen, Bertille Antoine, Dave Freeman, Ken Kasa, Erik Kimbrough, Chris Muris, and Simon Woodcock, all of whom have been generous with their time and knowledge when I needed them.

I have been fortunate to be in the company of excellent peers: Nick Dadson, Edouard Djeutem, Natt Hongdilokkul, Yang Li, Leanna Mitchell, Nick Kasimatis, Reza Sattari, Hanh Tong, Mike Tseng, Jing Wang, Xiao Yu and Keyi Zhang. Their friendship and support are indispensable in my journey as a Ph.D. student.

Finally, my thanks go to the administrative team of the graduate program—Simon Wookcock, Gwen Wild and Kathleen Vieira-Ribeiro—whose excellent work made my years at SFU a pleasant and memorable experience.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approval</td>
<td>ii</td>
</tr>
<tr>
<td>Ethics Statement</td>
<td>iii</td>
</tr>
<tr>
<td>Abstract</td>
<td>iv</td>
</tr>
<tr>
<td>Dedication</td>
<td>vi</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>vii</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>viii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xii</td>
</tr>
<tr>
<td>1 Cooperation, Competition, and Linguistic Diversity</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1 Motivation</td>
<td>1</td>
</tr>
<tr>
<td>1.1.2 Preview of the Results</td>
<td>2</td>
</tr>
<tr>
<td>1.1.3 Main Contributions</td>
<td>4</td>
</tr>
<tr>
<td>1.2 Related Literature</td>
<td>5</td>
</tr>
<tr>
<td>1.2.1 Linguistic Diversity and Its Correlates</td>
<td>5</td>
</tr>
<tr>
<td>1.2.2 Existing Explanations of Linguistic Diversity</td>
<td>6</td>
</tr>
<tr>
<td>1.2.3 Strategic Incentives as a Determinant of Linguistic Diversity</td>
<td>7</td>
</tr>
<tr>
<td>1.3 Model</td>
<td>10</td>
</tr>
<tr>
<td>1.3.1 Geography</td>
<td>10</td>
</tr>
<tr>
<td>1.3.2 Language</td>
<td>11</td>
</tr>
<tr>
<td>1.3.3 Strategic Interactions</td>
<td>14</td>
</tr>
<tr>
<td>1.4 Short Run Results</td>
<td>17</td>
</tr>
<tr>
<td>1.5 Long Run</td>
<td>24</td>
</tr>
<tr>
<td>1.6 Comparative Statics</td>
<td>28</td>
</tr>
<tr>
<td>1.7 Discussion</td>
<td>29</td>
</tr>
</tbody>
</table>
2 Strategic Incentives and Language Change 31

2.1 Introduction .................................................. 31
2.2 The Chen-Mitchell Model .................................. 31
2.3 Experimental Design ....................................... 34
  2.3.1 Subjects and Treatments ............................. 35
  2.3.2 Timeline of a Session ................................. 35
  2.3.3 Dialects and Linguistic Similarity .................. 36
  2.3.4 Period Game (Phase 2 Interaction) ................. 38
  2.3.5 Capturing the Evolution of Languages ............. 39
  2.3.6 Payment ............................................... 42
2.4 Results .................................................... 42
2.5 Related Literature ......................................... 46
2.6 Conclusion ................................................ 46

3 The Evolutionary Foundation of the Preference for Surprise 48

3.1 Introduction ................................................ 48
3.2 Related Literature ......................................... 51
  3.2.1 Evolution as Fitness Maximization: The Principal-Agent Metaphor 51
  3.2.2 Preference for Surprise ............................... 52
3.3 Model ...................................................... 52
3.4 Optimal Utility Function for the Agent .................. 56
  3.4.1 Utility as Fitness: Agent’s Problem ................ 56
  3.4.2 First Best Solutions ................................. 57
  3.4.3 Agent’s Problem with Utility for Surprise ........ 58
  3.4.4 Evolutionary Optimality of the Utility for Surprise 59
3.5 Discussion .................................................. 61

Bibliography ...................................................... 62

Appendix A Supplementary Materials for Chapter 1 68
A.1 Justification for the Reduced Form Expected Payoffs ...... 68
  A.1.1 General Environment ................................. 68
  A.1.2 Cooperative Interaction .............................. 70
  A.1.3 Competitive Interaction .............................. 70
A.2 Proofs ...................................................... 70
  A.2.1 Proof of Lemma 1.2 ................................ 70
  A.2.2 Proof of Proposition 1.3 ............................ 71
  A.2.3 Proof of Proposition 1.4 ............................ 72
  A.2.4 Proof of Proposition 1.5 ............................ 72
A.3 Figures ...................................................... 75
List of Tables

Table 2.1 Regression Analysis ........................................... 45
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1</td>
<td>Neighborhood Illustration</td>
<td>11</td>
</tr>
<tr>
<td>Figure 1.2</td>
<td>Timing of events in a typical period $t$</td>
<td>17</td>
</tr>
<tr>
<td>Figure 1.3</td>
<td>Optimal $</td>
<td>N^*_i</td>
</tr>
<tr>
<td>Figure 1.4</td>
<td>Optimal $</td>
<td>E^*_i</td>
</tr>
<tr>
<td>Figure 2.1</td>
<td>Neighborhood Illustration</td>
<td>32</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>Timeline of a Session</td>
<td>36</td>
</tr>
<tr>
<td>Figure 2.3</td>
<td>Sample Message List for a Subject</td>
<td>37</td>
</tr>
<tr>
<td>Figure 2.4</td>
<td>Phase 1 Decision</td>
<td>38</td>
</tr>
<tr>
<td>Figure 2.5</td>
<td>Phase 2 Games</td>
<td>39</td>
</tr>
<tr>
<td>Figure 2.6</td>
<td>Phase 2 Decisions</td>
<td>40</td>
</tr>
<tr>
<td>Figure 2.7</td>
<td>Subject-Message-List Association before and after Rotation</td>
<td>41</td>
</tr>
<tr>
<td>Figure 2.8</td>
<td>End-of-Round Display</td>
<td>43</td>
</tr>
<tr>
<td>Figure 2.9</td>
<td>Evolution of Message Lists over Time</td>
<td>44</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Optimal number of signals as a function of $v'$ (or $v$)</td>
<td>58</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Comparison between $n^A$, $n^*$, and $\hat{n}$</td>
<td>61</td>
</tr>
</tbody>
</table>
Chapter 1

Cooperation, Competition, and Linguistic Diversity

Joint work with Leanna Mitchell

1.1 Introduction

1.1.1 Motivation

The *Ethnologue*, a common reference for language classification, documents 7,105 human languages that are currently spoken worldwide (Lewis et al., 2013). These languages are distributed very unevenly on earth. Papua New Guinea, for example, accounts for only 0.3% of the world’s land mass, yet it is home to 12% of all living languages. Australia, by contrast, accounts for 5.1% of the world’s land area, but hosts merely 3% of its languages. Further examples of regions with high linguistic diversity include sub-Saharan west Africa and south-central Mexico (Lewis, 2009). Examples of areas with low linguistic diversity include northern Asia, Australia, and Brazil. In previous studies, scholars have documented relationships between geography and linguistic diversity. The most notable among those

---

1 The *Ethnologue* uses the ISO 639-3 standard, which classifies languages using a three-letter coding system. The basic criteria that the *Ethnologue* uses to identify languages (as opposed to, for example, dialects) are (i) mutual intelligibility between speakers of variants of a language; and (ii) existence of a common literature or of a common ethnolinguistic identity. (Lewis et al., 2013) The criterion for including a language in the *Ethnologue* is that it must be “known to have living speakers who learned [it] by transmission from parent to child as the primary language of day-to-day communication”. (Lewis, 2009)

2 See Figure A.2 in Section A.3 for an illustration.

3 These figures are the authors’ calculation based on data from WolframAlpha (www.wolframalpha.com) and Lewis et al. (2013). Specifically, the world’s land area is $1.4894 \times 10^9 \text{ km}^2$, that of Papua New Guinea is $462,840 \text{ km}^2$, and of Australia, $7,618 \times 10^6 \text{ km}^2$. The number of languages currently in use is 836 for Papua New Guinea and 245 for Australia.

4 Following the precedent of Michalopoulos (2012), we use the term *linguistic diversity* to refer to the number of languages in a region. Other terms, such as “language diversity” (Nettle, 1998; Pagel, 2000), and “density of language groups” (Mace and Pagel, 1995), are also used to refer to similar though not necessarily identical concepts.
studies, from an economics perspective, is the one by Michalopoulos (2012), who finds that variance in land quality and in altitude are positively and significantly correlated with linguistic diversity.

Furthermore, human linguistic diversity changes over time. Pagel (2000) estimates that between 130,000 and 500,000 languages have been spoken on this planet, and that linguistic diversity peaked between 20,000 and 50,000 years ago. Despite this historical plurality, the number of languages is decreasing rapidly. Recent evidence suggests that we are on a trend towards the “hegemony” of a few dominant languages. It has been estimated that between 50% and 90% of the current languages will not survive to the next century (Hale et al., 1992; Austin and Sallabank, 2011), although there are also many examples of linguistic groups in the process of reviving their traditional languages (Bentahila and Davies, 1993).

A third puzzle is the diversity of trajectory between pairs of languages or dialects. A living language changes constantly, borrowing words from other languages, developing new ones, and abandoning others. Each language has many variants, or so-called dialects, over both space and time, such as Canadian English vs. Australian English, or Modern English vs. Middle English. Two languages can merge into one. When they merge within a generation or two the daughter language is referred to as a creole (Hall, 1966). A language may also split and diverge many times. All of the 1,200 modern Malayo-Polynesian languages, for example, are believed to have descended from a common linguistic ancestor (Gray et al., 2009). In the case of divergence, the rates at which pairs of languages diverge are not uniform, nor is the rate of divergence for a single pair over time. As illustrated by Figure A.3, studies on the Indo-European and Malayo-Polynesian language families reveal that, while pairs of languages in the former family have diverged at a more or less common rate, languages in the latter family exhibit high variation in their rates of divergence, sometimes up to a threefold difference (Swadesh, 1952; Kruskal et al., 1971; Pagel, 2000).

### 1.1.2 Preview of the Results

In this paper, we show how, over time, cooperative and competitive incentives determine linguistic diversity. High frequency of cooperation, such as trade, has a homogenizing effect on language, thus leading to low linguistic diversity. High frequency of competition such as warfare, on the other hand, raises the value of the privacy of information, thus leading to the generation of new linguistic expressions and, over time, high linguistic diversity.

In situations involving potential benefit from cooperation between autonomous groups, the parties’ ability to communicate is undoubtedly critical. Trade, exogamous marriage and political alliances all rely on clear communication. Perhaps less conventionally, communication is also critical in situations of conflict. Theoretical work on conflict suggests that information plays a critical role in determining the outcome of a hostile contention.
In particular, there are “incentives to misrepresent information [...] specifically, each party would like to appear tougher than they really are” (Garfinkel and Skaperdas, 2007). Information regarding the strength of a group and any plans they have regarding attack or defense is particularly important. Success is more likely in these situations when a party can communicate such information within the group while keeping it secret from their opponents. The invention of new terms that are understood only by the “insiders” of the group facilitates the control of information.\footnote{Group-specific vocabulary and other common elements of language (e.g. accents) may also serve the purpose of identifying insiders from “outsiders” of the group. Accurate identification of fellow members enables groups to communicate strategically sensitive information to its membership without accidentally revealing it to outsiders. Accents would be especially useful for the purpose of identifying members of groups that are closely related linguistically. We do not formally model differentiated language as enabling identification in this way, but this use of language is definitely consistent with our hypothesis that linguistic groups will choose to invent more new terms in high conflict settings. Language as an identifier is especially useful in groups large enough that not all members know each other personally.}

In our model, groups interact periodically in pairs, in one of three types of activities: cooperation, competition, and non-interaction (hereafter referred to as a “null game”). Each group speaks a dialect, and two dialects are considered the same language if they are sufficiently similar.\footnote{The definitions of dialect and language will be made more precise in Section 1.3.} The groups’ expected payoffs in a game are a function of the relationship between their dialect and their opponent’s dialect. In cooperation, group $i$’s payoff is increasing in the commonality between $i$ and $j$’s dialects. This captures the intuition that coordination is more likely to be successful when the two groups can communicate in an accurate and inexpensive manner.

During conflict, on the other hand, $i$’s expected payoff increases in how much it understands the opponent $j$’s dialect, while it decreases in the level of $j$’s comprehension of $i$’s dialect. Intuitively, when $i$ and $j$ are in an antagonistic relationship, the more information $i$ has about $j$, the more advantaged $i$ would be in that relationship. Since language is the most common medium of intra-group communication, a better understanding of $j$’s dialect leads to a higher chance of obtaining $j$’s private information. By the same token, the less $j$ understands $i$’s dialect, the better $i$ would be able to protect its own information.\footnote{There is an additional, complementary benefit to a group of speaking a differentiated dialect, which we do not discuss at length or model explicitly. Differentiated language—including both vocabulary and accent—can help group members identify each other, which in turn facilitates the in-group distribution of sensitive information.}

Lastly, if two groups do not interact in a period, then their payoffs are unaffected by the relationship between their dialects.

In the model, each group interacts with every other group exactly once per period, and the nature of each interaction (i.e. cooperative, competitive, or null) is determined by a random draw for every pair from among the three games. For a pair $i, j$, the probability of drawing a cooperative game is $p_{ij}$, the probability of drawing a competitive game is $q_{ij}$, and the probability of drawing a null game is $1 - p_{ij} - q_{ij}$. The null game is more likely to occur when groups are farther apart geographically.
We also assume that the groups are myopic, so that there is no need to consider intertemporal tradeoffs. To maximize expected payoff, group $i$ can make a costly effort to change its dialect at the beginning of a period by (i) learning parts of the dialect(s) spoken by some other group(s), and/or (ii) inventing novel linguistic expressions that are only understood by members of group $i$. The benefit of learning from an existing dialect is twofold: first, it boosts $i$'s chance to succeed in a cooperative interaction with speakers of that dialect; second, it increases $i$'s competitive edge in conflict against speakers of that dialect, because $i$ would be more likely to decipher $j$'s intra-group communication. Inventing new expressions for intra-group communication, on the other hand, improves $i$'s odds of winning a conflict against all other groups, as the new expressions make it harder for all other groups to acquire information on $i$.

The main results we derive in this paper fall into three categories: (i) optimal linguistic change in a period game; (ii) the existence and characterization of a steady state in the infinitely repeated game, and convergence towards this state; and (iii) the steady state number of languages in a region and its comparative statics.

In any given period, it will always be optimal for a group to invent some new linguistic expressions, assuming that competitive games happen with positive probability. With regard to learning existing dialects, we show that if the probability of a non-trivial interaction with a neighbor drops sufficiently fast with an increase in geographical distance, then it is optimal for a group to learn from the dialect spoken by its closest neighbor(s) first.

In the dynamic setting, where the period game is played repeatedly, we establish the existence and uniqueness of a steady state in which the set of groups speaking each language remains unchanged over time. While languages themselves change in the steady state, the set of groups speaking any particular language does not. This in turn generates a unique steady state number of languages. If the initial linguistic composition is of a particular symmetric structure, we show that the number of languages in the region converges to that in the steady state. Lastly, we show that the steady state number of languages is weakly decreasing in the relative probability of cooperative vs. competitive interactions.

### 1.1.3 Main Contributions

Our paper makes three main contributions to the economics literature. First and foremost, we develop a formal economic model to explain how strategic incentives can induce language change and therefore linguistic diversity. Economic theorists have a long, albeit sporadic, interest in language. Marschak (1965) argued that the adaptability of languages to the environment in which they are used helps shape the features that they possess. Blume et al. (1993), Wärneryd (1993), and Robson (1990), among others, studied how messages in a language become associated with meaning in cooperative interactions with communication. Rubinstein (2000) characterized an “optimal” language based on criteria such as the ability...
of a language to identify objects and the ease by which the language is learned.\footnote{See Lipman (2003) for a good review.} More recently, \textit{Blume and Board (2013b)} explored the implications of language competence, and the knowledge thereof, on the efficiency of communication in common interest games. We contribute to this line of research by focusing on yet another aspect of language—the plurality of languages—and providing theoretical arguments for why such a phenomenon may be observed.

Second, our paper highlights the possibility that conflict generates a force that works against linguistic convergence. This is similar in spirit to the thesis of \textit{Blume and Board (2013a)}, that conflict of interest among speakers would lead people to favor a language that is vague even though more precise alternatives are available. Our model formalizes and extends the argument given by \textit{Pagel (2012)}, who hypothesizes that the multitude of languages exists to prevent people from understanding each other. In addition to conflict as a diverging influence on languages, our model also incorporates the traditional, more intuitive idea that cooperation fosters linguistic homogeneity. As a result, our model is more complete and produces a richer set of predictions regarding global linguistic diversity and linguistic change.

Third, this paper complements the existing literature on trade and conflict by explaining how trade and conflict may affect the diversity of languages and ethnicity. It is common in the economic literature to regard language as a facilitator of trade (\textit{Lazear, 1999}) and as a cause of conflict (\textit{Esteban et al., 2012}). Our theory, in contrast, suggests an alternate channel of causality.

The rest of the paper is organized as follows. \textit{Section 1.2} provides an overview of the stylized facts regarding linguistic diversity, current explanations, and evidence supporting our hypothesis. \textit{Section 1.3} introduces our model in a formal setting. In \textit{Section 1.4}, we derive the optimal language change in a typical period. \textit{Section 1.5} introduces the dynamic setting, establishes the existence, uniqueness and characterization of the steady state, and convergence thereto. The comparative statics of the steady state are given in \textit{Section 1.6}. \textit{Section 1.7} concludes.

\section{1.2 Related Literature}

\subsection{1.2.1 Linguistic Diversity and Its Correlates}

Scholars in several disciplines have documented empirical correlations between linguistic diversity and various ecological and geographic factors. Using historical data on native North American populations at the time of European contact, \textit{Mace and Pagel (1995)} find a significant positive correlation between linguistic diversity and the diversity of mammal species, both of which, in turn, exhibit a pronounced negative latitudinal gradient. On
a global scale, Harmon (1996) documents a positive correlation between linguistic and biological diversity. Also on a global scale, Nettle (1998) identifies climate as a key factor influencing global language distribution. In particular, he observes that areas with low rainfall and short growing season sustain fewer languages. Michalopoulos (2012) shows that contemporary linguistic diversity is related to geographic heterogeneity. He finds that variation in land quality (suitability for agriculture) and in elevation are both positively and significantly correlated with the number of languages in a particular region. Both Michalopoulos (2012) and Nettle (1998) find that average precipitation is significantly positively correlated with linguistic diversity.

There is a large and active literature in economics establishing a correlation between ethnolinguistic diversity and conflict. In this literature, measures of linguistic diversity or distance are used as proxies for differences in preferences over public goods (Esteban et al., 2012; Fearon, 2003; Desmet et al., 2012). The authors generally use linguistic data for the independent variable of interest because it is more available than data on genetic distance or on differences in preferences.

Additionally, there is a literature on the correlation between linguistic difference and trade or settlement patterns. Falck et al. (2012) find that similarity of dialect affects settlement decisions in Germany. Anderson and Van Wincoop (2004) review the literature showing that communication barriers increase the cost of trade between countries.

### 1.2.2 Existing Explanations of Linguistic Diversity

Several theories have been put forth to account for the empirical relationship between linguistic diversity and its environmental covariates. Common among these theories is a combination of isolation and drift. Linguistic divergence begins when populations of a linguistic community become isolated from one another. Drift in linguistics is a phenomenon analogous to mutation and drift in population genetics; it is random and unconscious change that can occur in language. When isolated, languages become dissimilar over time due to drift. Once the dissimilarity passes some threshold of mutual unintelligibility, those dialects would be considered different languages.

Nettle (1998) and Pagel (2000) argue that regions with favorable geographic conditions, e.g. those that are conducive to steady food supply, tend to sustain small, self-sufficient groups which seldom interact with each other. Consequently, populations in those regions are separated into various linguistic communities. In unproductive territories, on the other hand, survival demands that people cooperate on a large scale, and so language there will likely be uniform.

Boyd and Richerson (1988) point out another channel through which geography may increase isolation and therefore linguistic diversity. They maintain that high geographic heterogeneity increases migration costs, and thus facilitates isolation of different groups of
population. Then, due to the force of linguistic drift, the languages of these geographically separate communities become dissimilar over time. Michalopoulos (2012) also theorizes that geographically heterogeneous regions foster location-specific human capital that is not easily transferable to a different environment. As a result, possession of location-specific human capital contributes to the immobility of language groups. Via a mechanism similar to the one in Boyd and Richerson (1988), then, this immobility leads to a higher linguistic diversity within the region. Michalopoulos (2012) also notes that, as geographically uniform territories are easier to conquer and invasion has a homogenizing effect on language and culture, one would expect to see a positive correlation between linguistic diversity and geographic heterogeneity within a region.

Closest to our line of thinking is a short article by Pagel (2012). Pagel speculates that conflict over resources must be part of the reason why there are so many languages. He further mentions that humans are highly attuned to differences in speech, which helps them to identify the social group(s) to which individuals belong.

1.2.3 Strategic Incentives as a Determinant of Linguistic Diversity

We argue that changes in linguistic diversity are, at least partially, influenced by strategic incentives. Over many generations, strategically induced changes may accumulate and generate new languages. We propose a causal channel for such linguistic change: a region’s geographic make-up (e.g. climate, terrain, soil, vegetation, natural resources, bodies of water, etc.) affects the relative probability of cooperative vs. competitive interaction. This relative probability in turn determines the strategic incentives that drive linguistic change. Our theory therefore differs from the ones reviewed in the previous subsection by highlighting strategic considerations as an important causal factor.9

The relationship between geography and conflict has received much attention from geographers and political scientists. Geographic variables that are found to affect regional levels of conflict include general resource abundance (Kratochwil, 1986; Fairhead, 2000; Le Billon, 2001; Renner, 2002; Smillie and Forskningsstiftelsen, 2002); high value resources such as gems, fuel, or narcotics (De Soysa, 2002; Fearon and Laitin, 2003; Buhaug and Lujala, 2005; Buhaug et al., 2009); presence of mountains (Fearon and Laitin, 2003; Buhaug and Rød, 2006); forest cover (de Rouen and Sobek, 2004); and whether there is a rainy season (Buhaug and Lujala, 2005).10

---

9While we agree that some linguistic change is random, i.e. that there is linguistic drift, we have not included this force in the current version of the model in the interest of simplicity. Degree of isolation, on the other hand, does play a role.

10To our knowledge, no empirical study has been done to examine the effect of local variation in territorial quality on local conflict. We expect that if such a study were done, it would find a positive relationship between these two variables. Such evidence would link our theory to the empirical findings of Michalopoulos (2012); i.e. that variance in land quality is an important determinant of linguistic diversity.
There are several ways in which regional geography affects potential gains from trade. Factors such as climate and soil differ across sites in a region, generating comparative advantages in the production of particular goods. The magnitudes of these comparative advantages, in turn, affect the potential for gains from trade. Secondly, geography affects the transaction costs of trading. The costs of travel, and therefore the costs of trade, depend on geographic variables including ruggedness of terrain, availability of freshwater, and the location of water bodies.

There are abundant examples of cooperative incentives resulting in linguistic change. During the colonial period, many linguistic groups sought to develop new trading relationships with each other. As a result, hundreds of pidgin languages emerged (Hall, 1966). Pidgins are informal linguistic hybrids that arise when two groups of formerly isolated peoples need to communicate extensively with one other. In subsequent generations, a pidgin may become a creole, which has a more formal set of vocabulary and grammatical rules and serves as a first language for many people. Some examples of creoles are the Chavacano language in the Philippines, Krio in Sierra Leone, and Tok Pisin in Papua New Guinea (Hall, 1966). Cooperative incentives can also cause linguistic change on a much smaller scale. The English words “ranch”, “alligator”, and “barbeque”, for example, were borrowed from Spanish with only slight modification (Simpson et al., 1989).

There is a great deal of anecdotal evidence that conflict can lead to the invention of new words in the short run, and to linguistic divergence in the long run. As an example of the first case, Kulick (1992) reports that a group of Selepet speakers in Papua New Guinea changed their word for “no” from bia to bune to be distinct from a neighboring group. He also observed another Papua New Guinean linguistic group switching all of their masculine and feminine words—the words for mother and father, for example, switched in meaning. Kulick states that “people everywhere use language to monitor who is a member of their tribe.” It is very difficult to acquire the ability to speak a non-native language, \( \lambda \), perfectly. If a person manages to do so, then she is very likely to have a close relationship with members of the group whose native language is \( \lambda \).

Most social groups innovate linguistically, as exemplified by nicknames, slang, and “inside” jokes. Such terms can be useful in situations of inter-group conflict: for example, teenagers vs. parents, police vs. criminals, high school A vs. high school B, or government vs. insurgents. In these cases, a group can gain a strategic advantage by using newly invented expressions to identify group membership and communicate without revealing information to outsiders. A well documented and extensive case of this type of linguistic invention began in East London during the 1840s (Partridge et al., 2008). Speakers of the English dialect Cockney began generating “rhyming slang” which is relatively easy for insiders to learn but incomprehensible to outsiders. Various hypotheses have been proposed to explain the appearance of Cockney rhyming slang. These include linguistic accident, a form of amusement, to confuse the authorities, or, as Hotten (1859) believed, by
“chaunters and patterers”, i.e. street traders, possibly to assist with collusion. Developing some group-specific language naturally facilitates the control of information, and thereby generates a competitive advantage. In the case of Cockney rhyming slang, the police and customers were both engaged in competitive interactions with speakers of Cockney in the time period during which the slang emerged.

On a larger scale, there are many examples of distinct languages that have emerged in high conflict situations. “Thieves’ Cant”, a term that refers collectively to dozens of dialects of English which arose during the 17th century in Great Britain, was primarily spoken by criminal groups (Coleman, 2008). The Middle East (Goldschmidt and Davidson, 1991), the northern border of Italy (Kaplan, 2000), Nigeria (Osaghae and Suberu, 2005; Suberu, 2001) and Papua New Guinea (Johnson and Earle, 2000) are all examples of regions with both long histories of conflict and according to the Ethnologue, very high current linguistic diversity.

There are also many cases where groups in a competitive environment invest resources in maintaining a distinct language. An important category is the revival of indigenous languages, such as Halkomelem in Canada (Galloway, 2007), Welsh in Wales (Aitchison and Carter, 2000), Xibe in China (Jang et al., 2011), and Basque in Spain (Gardner et al., 2000). Minority groups will often fiercely protect their languages, as is the case for the Quebecois in Canada (d’Anglejan, 1984) and the Catalan in Spain (Roller, 2002). When facing conflict with sub-groups, a national government will sometimes try to eliminate the sub-group’s language or dialect. Some examples of such elimination include the mandatory residential school system for aboriginals in Canada in the 19th and 20th centuries (Milloy, 1999), and the establishment of the L’Académie française in 1635 with the goal of standardizing the French language (Rickard, 1989). Both policies arose during periods of internal conflict.

If drift is the only force affecting rates of linguistic divergence, these rates should be roughly constant over time. The two graphs in Figure A.3 from Pagel (2000), show that rates of divergence are instead highly variable over time. Rates of divergence are particularly high initially, when a pair of groups undergoes a linguistic split, and then decrease quickly, finally flattening out. Note that the scale on the y axis is logarithmic, so the change over time in the rate of divergence is very pronounced. We hypothesize that a single language would split into two during a period of conflict, when each faction found it beneficial to differentiate their language. As the pair of languages became less mutually intelligible, the incentives to differentiate them further would diminish. Furthermore, any change in underlying circumstances that happened to increase the relative frequency of cooperation vs. conflict would further decrease incentives to differentiate. In both graphs, the rate of divergence between any pair of languages clearly starts high, decreases quickly initially and then tapers off.

Furthermore, rates of divergence should be roughly the same for pairs of groups that are completely isolated from one another. An interesting fact about Figure A.3 is that
there is a large range of divergence rates among Polynesian languages, and a somewhat lesser range of divergence rates among pairs of Indo-European languages. The higher variance in divergence rates among the Polynesian languages is consistent with strategic incentives affecting these rates. The Polynesian islands host a very large variety of political relationships, with some being intensely and chronically violent, and others almost perfectly peaceful (Younger, 2008).

Last but not least, as Chapter 2 reports, in a randomized and controlled experimental setting, linguistic diversification is observed when subjects interact in competitive zero-sum games and linguistic convergence emerged when subjects interact in coordination games. These findings provide strong support for the theoretical predictions made in this paper.

1.3 Model

In our model, groups residing in a region interact myopically in every period. There are three types of interactions—cooperative, competitive, and null—which we use to model cooperation, competition, and no interactions, respectively. The probability of each type of interaction occurring is determined by the geographic environment, which is captured by a fixed, exogenous, region-wide parameter, and the distance between the interacting groups. The payoffs of the interactions depend on the groups’ understanding of each other’s dialects. Therefore, at the beginning of a period, the groups, anticipating the type of interactions they may be involved in, choose to update their existing dialect by either learning words from other groups or inventing new linguistic expressions. These linguistic changes, as well as the region’s linguistic composition, in turn determine the groups’ payoffs in the periodic interactions.11 The geographic parameters are constant over time, while the linguistic variables, to be introduced in Section 1.3.2, are in general time varying. We use a superscript $t$ on the linguistic variables to index time. However, in this and the next section, where it is clear from the context that the variable refers to one in a particular period $t$, we may suppress this time index.

1.3.1 Geography

A region of size normalized to 1 is populated by groups from the set $\mathcal{G} = \{1, \ldots, G\}$, where $G = 2^n$, $n \in \mathbb{N} \cup \{0\}$. Each group $i \in \mathcal{G}$ controls a site in the region. To model distances between sites, we introduce a neighborhood structure analogous to a binary tree, as depicted in Figure 1.1. A degree 1 neighborhood is inhabited by two groups; they are each other’s closest neighbors. A degree 2 neighborhood consists of two degree 1 neighborhoods, or four groups; and a degree $k$ neighborhood consists of two degree $k - 1$ neighborhoods. Hence, there are $2^k$ groups living in a degree $k$ neighborhood.

11A region’s linguistic composition is a description of which group speaks which language, and how much each group understands of the other groups’ dialects. The term is formally defined in Definition 1.5.
The degrees of neighborhood between two groups can be interpreted as the number of natural geographic barriers between them. For example, we can think of Sub-region 1 in Figure 1.1 as a common meeting ground for groups 1 and 2 (living in Sites 1 and 2), and the Region is the common place for groups 1 or 2 to meet groups 3 or 4. To get to a Sub-region, a group needs to travel a short distance, getting across a river, for example. On the other hand, to get to the Region, a group has to travel a longer distance, for instance, overcoming a big mountain range. For any two groups $i$ and $j$, let $d_{ij} \in \{0, 1, \ldots, D\}$ denote the smallest degree neighborhood $i$ and $j$ share. Then $d_{ij}$ is a proxy for the traveling time from the site of group $i$ to that of group $j$. A higher $d_{ij}$ indicates a farther distance between groups $i$ and $j$.

As will be clear in Section 1.3.3, the higher a pair’s neighborhood degree, the less likely they will interact in a given period. $D$ is the highest degree of neighborhood possible in the region (in Figure 1.1, for example, $D = 2$). Thus, $D = \log_2 G$, or equivalently, $G = 2^D$.

### 1.3.2 Language

For simplicity, we model a language as a list of sound-meaning pairs, which we call linguistic elements. Let $L$ be the set of all possible linguistic elements. For technical convenience, we

---

12. Trivially, $d_{ij} = d_{ji}$. Since any group is a degree 0 neighbor with itself, $d_{ii} = 0 \iff i = j$.

13. The binary tree neighborhood structure is a highly stylized way to model distance between groups. We choose this approach mainly for its tractability, rather than its resemblance of any geographic structure in the real world. This binary tree neighborhood structure allows us to use a single parameter, $d$, to keep track of both the distance between any two groups and the number of groups within a particular degree of neighborhood ($2^d$). The latter feature is useful in proving Lemma 1.1, and subsequently establishing the existence of a steady state.

14. That is, $L$ contains all possible sound-meaning pairs. For example, “perro”-dog, “haha”-dog, “cat”-dog and “dog”-dog would be four of the elements of $L$. Note also that $L$ is not time dependent.
assume that \( \mathcal{L} = \mathbb{R} \).\(^{15}\) Let \( \mathcal{B} \) be the Borel \( \sigma \)-algebra on \( \mathbb{R} \), and \( \cdot \) be the Lebesgue measure on \( (\mathcal{L}, \mathcal{B}) \). Therefore, \( (\mathcal{L}, \mathcal{B}, \cdot) \) forms a measure space.

**Definition 1.1.** A dialect spoken by group \( i \) at the beginning of period \( t \) is a non-empty, measurable (with respect to \( \cdot \)) subset \( L^t_i \subset \mathcal{L} \) with a finite measure \( |L^t_i| < \infty \). Let \( L^t_0 = \mathcal{L} \setminus \cup_{k \in \mathcal{G}} L^t_k \) denote the set of linguistic elements that are not used in any dialect.

Group \( i \) can choose to enrich its dialect in two ways: (i) adopting linguistic elements from other groups’ dialects, i.e. from \( \{ L^t_j \setminus L^t_i : j \neq i \} \), or (ii) acquiring elements from \( L^t_0 \), which are not in use by any group.

Denote by \( E^t_i \) the set of linguistic elements \( i \) learns from other groups’ dialects. We write

\[
E^t_i = \bigcup_{k \in \mathcal{G} \setminus \{i\}} E^t_{ik}
\]

where \( E^t_{ik} \subseteq L^t_k \setminus L^t_i \) is the (possibly empty) set of linguistic elements adopted from \( L^t_k \) by \( i \). Assume that each \( E^t_{ik} \) is measurable. A non-empty \( E^t_{ik} \) indicates that group \( i \) is adopting/learning part of other groups’ vocabulary. This captures, for example, the adoption of English word by Spanish, and vice versa, as discussed in Section 1.2.3.

Let \( N^t_i \subseteq L^t_0 \) denote the set of linguistic elements \( i \) acquires from \( L^t_0 \). Again, each \( N^t_i \) is measurable.\(^{16}\) We further assume that groups acquire elements from \( L^t_0 \) in an uncoordinated fashion, and therefore that \( N^t_i \) and \( N^t_j \) are disjoint (up to a subset of measure zero) for any \( i \) and \( j \).\(^{17}\) The latter assumption is plausible, in that \( L^t_0 \) is of an infinite measure while each \( N^t_i \) has a finite measure, and so it is unlikely that any two groups would adopt the same subset of elements from \( L^t_0 \) when their actions are uncoordinated. This type of learning can be thought of as people in group \( i \) inventing new expressions for their own dialect. Thus, a non-empty \( N^t_i \) corresponds to the examples of “code” words developed by groups such as criminal organizations, the police, the Selepet, etc. when facing conflict. The emergence of Cockney rhyming slang and the revival of Hebrew are examples of very large \( N^t_i \).

We consider several dialects to be variants of the same language if they are sufficiently similar, as follows:

**Definition 1.2.** A subset \( \mathcal{G}^t \subseteq \mathcal{G} \) of groups speak the same language \( \lambda^t \) at the end of period \( t \) if \( L^t_i \cup E^t_i = L^t_j \cup E^t_j \) up to a subset of measure zero for all \( i, j \in \mathcal{G}^t \). The language \( \lambda^t \) is defined as the set \( \lambda^t = \bigcup_{k \in \mathcal{G}^t \setminus \{0\}} (L^t_k \cup E^t_k \cup N^t_k) \).

\(^{15}\)We do not require a notion of closeness of any two linguistic elements, however. Using an alternative assumption—letting \( \mathcal{L} \) be countable—will not change our results qualitatively. It is just awkward to work with a discrete set.

\(^{16}\)In the remainder of the paper, whenever we refer to a subset of \( \mathcal{L} \), we assume that it is measurable, unless otherwise noted.

\(^{17}\)Note also that \( L^t_i, E^t_i \) and \( N^t_i \) are disjoint pairwise. The names of the sets are chosen to signify their properties: \( E \) stands for “existing”, so elements in \( E^t_i \) are chosen from existing dialect (other than \( L^t_i \)); and \( N \) stands for “non-existing”, so that it contains elements that are not from an extant dialect.
Thus, two dialects are deemed the same language if they consist of the same set of linguistic elements, acquired either through inheritance from a previous generation (i.e. elements in the set $L^t_i$) or learning (i.e. elements in the set $E^t_i$). Some readers may wonder why we choose $L^t_i \cup E^t_i = L^t_j \cup E^t_j$, as opposed to $L^t_i = L^t_j$ (or $L^t_i \cup E^t_i \cup N^t_i = L^t_j \cup E^t_j \cup N^t_j$), as the criterion of similarity between $i$ and $j$’s dialects. Our choice is based on two reasons. First, languages evolve continually, and so it is unlikely for two dialects—in our model as well as in reality—to be identical at any point in time. Therefore, $L^t_i = L^t_j$ as a criterion of similarity, which basically requires $i$ and $j$’s dialects to be identical up to a subset of measure zero, would be too stringent. Second, in our model, especially the dynamic part in Section 1.5, groups learn and invent linguistic elements in every period. In this setting, $L^t_i \cup E^t_i = L^t_j \cup E^t_j$ is the most similar two dialects can get at the end of a given period. Hence Definition 1.2 is already using a very strict criterion, given the setup of our model.

Definition 1.2 is also consistent with the two criteria of language classification in the Ethnologue (see footnote 1): common ethnolinguistic source and mutual intelligibility. Every language has many variants, and these variants contain features specific to a particular linguistic group. The degree of similarity of pairs of variants is of a continuous nature. Above some arbitrary degree of dissimilarity, variants are referred to as distinct dialects, and above a second but higher arbitrary degree of dissimilarity they are counted as distinct languages. The elements in $N^t_i$ and $N^t_j$ represent exactly this aspect of dialects; their presence does not disqualify dialects from being counted as a single language.

Elements in $N$ are linguistic innovations, as discussed in Section 1.3.3, and often considered to be new “slang”. By counting English as a single language we declare Canadian English and American English, for example, to be the same language. These two variants of English, however, are by no means identical. What distinguishes them, differences in spelling, accent, language use, etc., exist partially due to the diverging force of competition. Hence, we consider it appropriate to ignore elements in $N^t_i$ and $N^t_j$ when judging whether groups $i$ and $j$ speak the same language.

Defining a language in this way allows us to determine the speakership of each language, and then count the number of languages in the region. This can be done for any period, and becomes particularly important in Section 1.5, when the model is extended to a dynamic setting where speakerships may change from period to period.

The cost of acquiring new linguistic elements within a period depends on the sizes of the sets acquired, $E^t_i$ and $N^t_i$, and is assumed to have the following (time independent)

---

18 In defining language and dialect, the Ethnologue makes the following comment: “Every language is characterized by variation within the speech community that uses it. Those varieties, in turn, are more or less divergent from one another. These divergent varieties are often referred to as dialects. They may be distinct enough to be considered separate languages or sufficiently similar to be considered merely characteristic of a particular geographic region or social grouping within the speech community.” (Lewis et al., 2013)

19 In this paper, we use the words “size” and “measure” interchangeably.
functional form:

\[ C(|E_i^t|, |N_i^t|) = c(|E_i^t|) + c(|N_i^t|) \]

where \( c : [0, \infty] \rightarrow \mathbb{R} \cup \{\infty\} \) is a strictly increasing and strictly convex function with \( c(0) = 0 \). Here we assume that there is no complementarity between the costs of the two modes of acquisition. Learning another language and inventing new expressions occur in very different contexts. Learning happens in an inter-group environment, in which multiple groups must spend time together; whereas inventing occurs in an intra-group context, where members of a single group spend time together inventing novel linguistic expressions. We assume the social and mental resources required for engaging in these activities are different enough to justify a zero cross-partial derivative.\(^{21}\)

To make our analysis more tractable, we assume that \( c \) is quadratic:

\[ c(|\cdot|) = \frac{1}{2} |\cdot|^2, \tag{1.1} \]

where \( |\cdot| \) is the measure of a set. Convexity represents the increasing marginal costs that are likely to occur with such activities.

### 1.3.3 Strategic Interactions

Our model is set in the context of small scale farming or foraging economies.\(^ {22}\) Pairs of groups in the region interact periodically in one of three activities: cooperation, competition, or non-interaction.\(^ {23}\)

A common example of a cooperative interaction is a situation in which there are potential gains from trade. The coastal First Nations of British Columbia, for example, often traded oolichan oil as far as 300 miles into the interior of the continent in exchange for commodities such as copper, flint, dried meat, and furs (Phinney et al., 2009). Accurate communication is very helpful in generating surplus in such transactions. Larger surpluses are possible if traders can easily and accurately communicate such information as when and where they will meet, and what the demand and supply are likely to be for each good.

Formally, we model a cooperative interaction as a pair of groups trying to coordinate on accomplishing some task using language. The expected payoff from a cooperation is increasing in the measure of the intersection \((L_i \cup E_{ij}) \cap (L_j \cup E_{ji})\).\(^ {24}\) The intuition is that, since \( i \) and \( j \) use language to coordinate their actions, the more functional linguistic elements

\(^{20}\)Since the Lebesgue measure maps \( B \) onto \([0, \infty] \), it follows that the image of \(|\cdot|\) is the entire non-negative half of the extended real line.

\(^{21}\)We do not believe that relaxing the additive separability assumption would change our results in a qualitative way. At the very least, we conjecture that our conclusions would still hold if we allow for a sufficiently small cross partial between the costs in the two modes of acquiring new linguistic elements.

\(^{22}\)See Section 1.7 for a discussion of this modeling choice.

\(^{23}\)This subsection deals with what happens in a typical period \( t \), so we suppress the superscript \( t \) on the linguistic variables.

\(^{24}\)See Section A.1 for an explanation.
they have in common, the more accurate their communication will be, and consequently the better the two groups will perform in the cooperative interaction. Hence, we define group $i$’s reduced form expected payoff from the cooperative interaction as follows:

$$ u_i(E_i, E_j) = |(L_i \cup E_{ij}) \cap (L_j \cup E_{ji})|. \tag{1.2} $$

Observe that $N_i$ and $N_j$ do not affect $u_i$. It would thus be equivalent to define $u_i$ as $|(L_i \cup E_{ij} \cup N_i) \cap (L_j \cup E_{ji} \cup N_j)|$, since $N_i \cap (L_j \cup E_{ji} \cup N_j) = \emptyset$.

A competitive game represents a situation where the two groups are antagonistic towards each other, for example when they are competing over the use of resources. We formally model the competitive game as a form of zero-sum game. The role of language in this context is to communicate within a group itself, for example for group $i$ to organize a show of strength, an attack, or a plan for defense. With positive probability, $i$’s within-group communication may be intercepted/overheard by someone from group $j$. If members of group $j$ know a large portion of the elements in $i$’s dialect, there is a high probability that an intercepted communication will be understood and used to group $i$’s disadvantage. To capture this formally, we define the (reduced form) expected payoff from the competitive game as follows:

$$ v_i(E_i, N_i, E_j, N_j) = \beta \left[ |(L_i \cup E_i \cup N_i) \setminus (L_j \cup E_j)| - |(L_j \cup E_j \cup N_j) \setminus (L_i \cup E_i)| \right]. \tag{1.3} $$

This is the difference between how much of $i$’s dialect is private from $j$, and how much of $j$’s dialect is private from $i$, weighted by $\beta$. The first term in the square brackets captures $i$’s ability to conceal information from $j$, and the second term reflects $j$’s ability to conceal information from $i$. Therefore, $i$’s expected payoff from the competitive game is increasing in the former and decreasing in the latter. The parameter $\beta$ represents the relative magnitude of competitive payoffs vs. cooperative payoffs. Note also that $v_i(\cdot) = -v_j(\cdot)$.

The null game can be interpreted as a scenario in which $i$ and $j$ do not interact. When a pair of groups play a null game, each gets a payoff of zero with certainty.

---

25We say that a linguistic element $\ell \in \mathcal{L}$ is functional for the purpose of cooperation between $i$ and $j$ if and only if $\ell \in L_i \cup L_j$. That is, two groups can use a linguistic element to enhance cooperation only if at least one of them is already fluent in using that element. This is plausible because miscommunication is more likely when two groups try to communicate using an element from a third dialect that both have recently learned. The purpose of this assumption is to simplify the characterization of a group’s optimal strategy $E^\ast_i$. The alternative is to let the payoff from a cooperative interaction depend on the measure of $(L_i \cup E_i) \cap (L_j \cup E_j)$. However, if non-functional elements, particularly those in $E_i \cap E_j$, can also influence the payoff in a cooperative interaction, then it becomes complicated to solve for the measure of $E_i \cap E_j$, and hence that of $(L_i \cup E_i) \cap (L_j \cup E_j)$.

26See Section A.1 for a derivation of equations (1.2) and (1.3).

27See, for example, Dawson (1891). The real world counterparts to this game, however, need not involve actual violence.

28See Section A.1 for details.

29Notice that the functionality distinction is only relevant in a cooperative game. Thus, even if both $i$ and $j$ learn the same (measurable) subset of elements from a third group $k$, knowledge of that subset of elements does not affect the payoffs of a competitive game between $i$ and $j$. 

15
For any matched pair $i$ and $j$, the probability of a cooperative interaction occurring is $p_{ij}$, a competitive game $q_{ij}$, and a null game $1 - p_{ij} - q_{ij}$. These probabilities are related to the distance between $i$ and $j$, $d_{ij}$, and the geography of the region through a parameter $r$ which summarizes the region’s geographic conditions.\footnote{Our theory does not depend critically on the relationship between $r$ and any specific geographic variables, as long as it is determined jointly by a set of relevant geographic factors, and is roughly constant throughout the region.} Let the relative probability of the cooperative vs. competitive game in a given period be

$$\frac{p_{ij}}{q_{ij}} = r,$$  \hspace{1cm} (1.4)

where $r$ is exogenous and constant throughout the region. The fraction $p_{ij} / q_{ij}$ is therefore the same for all pairs $i, j \in \mathcal{G}$. Let $\pi : \{0, \ldots, D\} \rightarrow [0, 1]$ be a function that relates the geographic distance between $i$ and $j$, $d_{ij}$, to the probability that these two groups will interact either cooperatively or competitively, i.e. $p_{ij} + q_{ij}$. Then, the probability of $i$ and $j$ playing a null game, i.e. they do not interact in a period, is

$$1 - p_{ij} - q_{ij} = 1 - \pi(d_{ij}). \hspace{1cm} (1.5)$$

We assume that $\pi(\cdot)$ satisfies the following two properties:\footnote{An example would be $\pi(d) = \gamma 2^{-d(d-1)/2}$, where $\gamma = \pi(1) \leq 1$.}

$$\frac{\pi(d)}{\pi(d+1)} \geq 2^d, \quad \forall d \in \{0, \ldots, D\} \hspace{1cm} (1.6)$$

$$\pi(0) = 1. \hspace{1cm} (1.7)$$

Property (1.7) can be loosely interpreted as that a group always “interacts with itself”. This is a requirement mainly for technical purposes. Property (1.6) implies that $\pi(\cdot)$ is a decreasing function. Thus, the farther apart two groups are geographically (i.e. the larger $d_{ij}$ is), the less likely they will interact with each other. Furthermore, property (1.6) requires that the probability of interaction between any two groups drops sufficiently fast as the distance between them increases. Specifically, we assume that the probability of a group interacting with one degree $d$ neighbor is greater than the sum of the probabilities of interacting with all of its degree $d + 1$ neighbors.\footnote{Note that property (1.6) can be written as $\pi(d) \geq 2^d \pi(d + 1)$, where $2^d$ is the number of degree $d + 1$ neighbors of a group.} This assumption is necessary for the proof of Lemma 1.1, which shows that $i$ always prefers to learn, when possible, from a closer neighbor’s dialect than from a farther neighbor’s.\footnote{As will be clear in the next section, the marginal benefit of learning from a dialect is directly related to the probability of interacting with the group that speaks it. Property (1.6) basically ensures that learning a word only known by one degree $d$ neighbor is more useful than learning a word known by $2^d$ degree $d + 1$ neighbors. If property (1.6) does not hold, then it becomes extremely difficult to characterize an equilibrium.}

Together, (1.4) and (1.5) imply that
The timing of events within a period, illustrated by Figure 1.2, is as follows:

i) At the beginning of a period, each group observes \( \{ |L_j \setminus L_i| : i, j \in \mathcal{G} \} \). In other words, each group knows the measure of the set that exists to be learned from every other group’s dialect.

ii) \( \{ E_i \} \) are chosen simultaneously.

iii) \( \{ N_i \} \) are chosen simultaneously.

iv) A period game is then drawn for every possible pair of groups, according to the probabilities described in (1.5) and (1.8).

v) Lastly, all \( G(G - 1)/2 \) games are played, and payoffs are realized.

In Section 1.4, we derive a group’s optimal decisions, \( N_i^* \) and \( E_i^* \), for a typical period. In Section 1.5, we examine the long run implications of the short run results.

1.4 Short Run Results

Our first result establishes the optimal size of \( N_i \) for every group \( i \). We show that each \( N_i^* \), the optimal subset of linguistic elements acquired by group \( i \) from \( L_0 \), has the same measure.

in this problem, as equilibria would then depend on the initial composition of dialects—the pattern of their pairwise intersections.
Proposition 1.1 (Optimal size of $N_i$). For $i \in \mathcal{G}$, we have $|N_i^*| = |N^*|$, where

$$|N^*| = \frac{\beta}{1 + r} \sum_{k=1}^{D} 2^{k-1} \pi(k).$$

(1.9)

Proof. From (1.2) and (1.3), the ex ante expected payoff of group $i$ is

$$U_i(\cdot) = \sum_{j \neq i} p_{ij} \left[ |(L_i \cup E_{ij}) \cap (L_j \cup E_{ji})| - c(|E_i|) \right] + \sum_{j \neq i} \beta q_{ij} \left[ |(L_i \cup E_i \cup N_i) \setminus (L_j \cup E_j)| - |(L_j \cup E_j \cup N_j) \setminus (L_i \cup E_i)| \right] - c(|N_i|).$$

(1.10)

Observe that only the second line involves $N_i$. Thus group $i$’s maximization problem is

$$\max_{|N_i|} \sum_{j \neq i} \beta q_{ij} \left[ |(L_i \cup E_i \cup N_i) \setminus (L_j \cup E_j)| - |(L_j \cup E_j \cup N_j) \setminus (L_i \cup E_i)| \right] - c(|N_i|).$$

Since benefit is linear and cost is strictly convex in $|N_i|$, and $c'(0) = 0$, there exists a unique maximum. The maximum is given by the first order condition where marginal cost is equal to marginal benefit.\textsuperscript{34} Notice that marginal benefit of $|N_i|$ is constant, described by

$$\beta \sum_{j \neq i} q_{ij} = \beta \sum_{j \neq i} \frac{1}{1 + r} \pi(d_{ij}) = \frac{\beta}{1 + r} \sum_{k=1}^{D} 2^{k-1} \pi(k),$$

where the first equality follows from equation (1.8) and the last equality follows from the fact $i$ has $2^{k-1}$ degree $k$ neighbors. The marginal cost of acquiring elements is just $|N^*|$, according to (1.1). Therefore condition (1.9) is precisely the first order condition. Since the problem is symmetric for every $i$, it follows that (1.9) holds for all $i \in \mathcal{G}$. \hfill \qed

The intuition for this result is the standard marginal analysis in economics, as depicted in Figure 1.3. Acquiring linguistic elements from $L_0$ means inventing new expressions that no other groups but $i$ can understand. Doing so enhances $i$’s ability to keep secrets from all other groups, which in turn raises $i$’s expected payoff in all of its competitive games. The marginal benefit is therefore the sum of the probabilities of playing a competitive game with each other group in the region, weighted by the relative magnitude of competitive payoffs to cooperative ones, $\beta$. At the (interior) optimum, the marginal benefit must equal marginal cost of inventing new linguistic expressions. Observe that $|N^*|$ depends only on $r, \beta$ and $D$, which are all exogenous and constant parameters and constant over time. Therefore, $|N^*|$ is constant over time as well.\textsuperscript{34}

\textsuperscript{34}Linear benefit and strictly convex cost ensures that the second order condition holds as well.
To obtain a similar characterization for $|E^*_i|$, we need to put some structure on the initial set of dialects, $\{L_i\}_{i \in \mathcal{G}}$. Specifically, we restrict attention to sets of dialects that are localized, as defined below.

**Definition 1.3 (Localization).** The set of dialects $\{L_i\}_{i \in \mathcal{G}}$ is *localized* if the following condition is satisfied:

$$\ell \in L_i \setminus L_j \implies \ell \notin L_k, \quad \forall i, j, k \in \mathcal{G} \text{ such that } d_{ij} < d_{ik}. \quad (1.11)$$

Localization means that if there is an element that a group $j$ can learn from $L_i$, then that element cannot also be in the dialect of any of $i$’s and $j$’s mutually farther neighbors.\(^{35}\) In other words, localization requires that whenever $L_i$ and a distant neighbor’s dialect $L_k$ have a common element, then all the dialects of $i$’s closer neighbors must also have that element.

Localization captures the intuition that it is very improbable for two distant languages to independently develop the same word that has the same meaning. Evidence from historical linguistics suggests that the elements of pre-colonial languages at least roughly satisfy the localization property. When a linguistic group splits into two, the speakers of the two new dialects naturally tend to live close together. Studies of the Indo-European languages, for example, found that languages with a more recent common linguistic ancestor—e.g. Spanish and Portuguese, both of which share many common features of Latin, their “parent language”—are also geographically close to each other (Finegan, 2008).

According to Pagel et al. (2007), “[L]anguages, like species, evolve by a process of descent with modification”. Linguistic divergences are also commonly represented as trees, similar to those relating biological species (Kruskal et al., 1971; Pagel, 2000; Esteban et al., 2012).\(^{35}\)

\(^{35}\)It is worth noting that both $i$ and $j$ are equally distant from $k$ when $d_{ij} < d_{ik}$. In Figure 1.1, suppose $i, j, k$ are Sites 1, 2, and 3, respectively. Sites 1 and 2 are degree 1 neighbors with $d_{12} = d_{21} = 1$, and both are degree 2 neighbors of Site 3 with $d_{13} = d_{23} = 2$. Thus, $d_{12} < d_{13}$ implies that $d_{21} < d_{23}$.
Suppose that \( i \) and \( j \) have a more recent common linguistic ancestor than \( i, j, \) and \( k. \) An element that satisfies any one of the three following conditions satisfies the localization property: (i) all three retain the element from a common ancestor, (ii) an ancestor of both \( i \) and \( j \) split from \( k \) and then acquired it before \( i \) and \( j \) split from each other (iii) any element acquired by a single group after all three had split.

There is also a theoretically appealing reason for our focus on symmetric and localized sets of dialects. As we show later in Lemma 1.2, if the region begins with a set of symmetric and localized dialects, then the set of dialects will always be symmetric and localized. If all groups in a neighborhood of some degree \( d \in \{0, 1, \ldots, D\} \) speak the same language, then the set of dialects is localized. Localization, however, holds for a much more general set of dialects.

Assuming localization enables us to establish the order in which a group \( i \) learns from existing dialects. This order is closely connected to the magnitude of the marginal benefit of learning a subset of another dialect. Let \( MB_i(d_{ij}, s) = 2^{d_{ij} - s}(p_{ij} + \beta q_{ij}) \) denote the marginal benefit that \( i \) receives from learning a subset of elements that is shared by \( 2^{d_{ij} - s} \) neighbors of degree \( d_{ij}, \) where \( s \in \{1, \ldots, d_{ij}\}.\)

**Lemma 1.1 (Order of Learning).** Suppose the set of dialects \( \{L_i\}_i \) is localized. Then, \( MB_i(d_{ij}, s) \) is decreasing in \( d_{ij}. \) As a result, it is always optimal for \( i \) to learn all of \( \bigcup_j (L_j \setminus L_i) \) before learning anything from \( \bigcup_k (L_k \setminus L_i), \) where \( j, k \) are such that \( d_{ij} < d_{ik}. \)

**Proof.** From Definition 1.3 and equation (1.2), it follows that after learning all the elements of all the languages where \( d < d_{ij}, \) learning a subset of elements \( E_{ij} = L_j \setminus L_i \) is only going to affect \( i \)'s payoff when it is interacting with neighbor(s) of degree \( d_{ij}. \) According to (1.10) and (1.8), therefore, given an arbitrary \( d_{ij} \in \{1, \ldots, D\}, MB_i(d_{ij}, s) \) is at least

\[
p_{ij} + \beta q_{ij} = \frac{r + \beta}{1 + r} \pi(d_{ij}), \tag{1.12}
\]

as is the case when \( s = d_{ij}, \) and at most

\[
2^{d_{ij} - 1}(p_{ij} + \beta q_{ij}) = 2^{d_{ij} - 1}\left(\frac{r + \beta}{1 + r} \pi(d_{ij})\right), \tag{1.13}
\]

as is the case when \( s = 1. \) \( s = 1 \) is the case where all of \( i \)'s degree \( d_{ij} \) neighbors know the element, so it will be useful in games with any of them. Observe that (1.12) and (1.13) are the
same when \( d_{ij} = 1 \), because \( i \) only has one degree 1 neighbor. Observe also that the lowest marginal benefit of learning from a degree \( d_{ij} \) neighbor, i.e. \( MB_i(d_{ij}, d_{ij}) \), is higher than the highest marginal benefit of learning from a degree \( d_{ij} + 1 \) neighbor, i.e. \( MB_i(d_{ij} + 1, 1) \):

\[
\frac{r + \beta}{1 + r} \pi(d_{ij}) - 2^{(d_{ij}+1)-1} \left( \frac{r + \beta}{1 + r} \pi(d_{ij} + 1) \right) \geq 0, \quad \forall d_{ij} \in \{0, \ldots, D\}.
\]

The inequality follows from (1.6). This completes the proof.

Lemma 1.1 allows us to partition the set of \( i \)'s learnable elements, \( \bigcup_k (L_k \setminus L_i) \), according to the marginal benefit they confer, and hence the neighborhood degrees. Let

\[
P_i(d) = \begin{cases} 
\{ \bigcup_k (L_k \setminus (\bigcup_{k'} L_{k'})) : d_{ik} = d \text{ and } d_{ik'} < d \} & \text{if } d \in \{1, \ldots, D\} \\
\emptyset & \text{if } d = 0
\end{cases}
\]
describe the elements of this partition.

\( P_i(d) \) is the set of learnable linguistic elements in the dialects of \( i \)'s degree \( d \) neighbors, excluding the elements that are in the dialects of \( i \)'s closer neighbors, of degrees less than \( d \). Note that elements of \( i \)'s own language are never learnable to \( i \); since \( i \) is a degree 0 neighbor with itself, \( P_i(d) \cap L_i = \emptyset \) for all \( d \). Moreover, for any \( d \neq d' \), \( P_i(d) \) and \( P_i(d') \) are disjoint.

It follows, therefore, that \( |\bigcup_d P_i(d)| = \sum_d |P_i(d)| \). By Lemma 1.1, it also follows that, for any \( A \subseteq P_i(d) \) and \( A' \subseteq P_i(d') \) where \( |A| = |A'| \) and \( d < d' \), the marginal benefit of learning \( |A| \) is strictly greater than that of learning \( |A'| \). Moreover, each \( P_i(d) \) can be further partitioned into \( d \) cells, based on how many neighbors of degree \( d \) share those elements.

By the definition of \( MB_i(d_{ij}, s) \), the more neighbors sharing an element, the higher the marginal benefit that element confers. Therefore, assuming localization, we can order \( i \)'s set of learnable elements by their associated marginal benefits: first by neighborhood degree \( d \), and then by \( s \) within each \( P_i(d) \). Such an ordering, together with Lemma 1.1, implies that as \( |E_i| \) increases, marginal benefit decreases in a step-wise fashion.

**Proposition 1.2 (Optimal size of \( E_i^* \)).** Let the set of dialects \( \{L_i\}_i \) be localized. The optimal size of \( E_i^* \) is contained within the following interval:

\[
|E_i^*| \in \left[ \sum_{k=0}^{d^*} |P_i(k)|, \quad \sum_{k=0}^{d^*+1} |P_i(k)| \right],
\]

where \( d^* \in \{0, \ldots, D\} \) is determined by

\[
\frac{r + \beta}{1 + r} \pi(d^*) \geq \sum_{k=0}^{d^*} |P_i(k)| \quad \text{and} \quad \frac{r + \beta}{1 + r} \pi(d^* + 1) < \sum_{k=0}^{d^*+1} |P_i(k)|. \tag{1.14}
\]
Our dynamic results in Section 1.5 do not require that we know the exact value of $|E_i^*|$, only the degree of neighborhood $d^*$ within which $i$ chooses to learn all the remaining elements of its neighbors dialects.

Proof. First, observe that a unique $|E_i^*|$ exists. The size of acquired elements from existing dialects, $|E_i|$, takes a value from a compact set $\left[0, \sum_{k=0}^{D} |P_i(k)|\right]$, on which the objective function (1.10) is continuous. Hence a maximum exists. Since marginal benefit is weakly decreasing in $|E_i|$ while marginal cost is strictly increasing, uniqueness of $|E_i^*|$ is ensured. Figure 1.4 provides an illustration.

We can verify that $|E_i^*|$ is indeed bounded by the proposed interval through the first order condition. Notice that the benefit function is an increasing, piece-wise linear function that is not differentiable at a finite number of points, namely the points at which marginal benefit makes a discrete jump downwards. Nevertheless, for each value of $|E_i|$, there exists a set of (super-)derivatives for the benefit function, bounded by the left- and right-derivatives at $|E_i|$, and the set is a non-singleton at the points where the benefit function has a kink. The first order condition requires that at $|E_i^*|$, there exists a (super-)derivative of benefit function that is equal to the derivative of the cost function (the latter of which is uniquely defined at every $|E_i|$). Moreover, it has to be true that (i) at $|E_i| = \sum_{k=0}^{d^*} |P_i(k)|$, marginal benefit is weakly higher than marginal cost; and (ii) at $|E_i| = \sum_{k=0}^{d^*+1} |P_i(k)|$ marginal cost exceeds marginal benefit. But according to (1.14), $d^*$ is chosen such that these two conditions are simultaneously satisfied. The marginal benefit of learning the last element in $P_i(d^*)$ is described by (1.12); for otherwise $i$ would have learned it sooner. The marginal cost of learning this last element is $\sum_{k=0}^{d^*} |P_i(k)|$. Therefore, equation (1.14) ensures that $d^*$ is chosen such that the marginal benefit of learning the last element in $P_i(d^*)$ is higher than the marginal cost. However, the marginal cost exceeds the marginal benefit if $P_i(d^* + 1)$ is fully acquired.
We know \(|E_i| = \sum_{k=0}^{d^*} |P_i(k)|\) when everything up to and including \(P_i(d^*)\) is fully learned and \(|E_i| = \sum_{k=0}^{d^*+1} |P_i(k)|\) when everything up to and including \(P_i(d^*+1)\) is fully learned. Therefore, \(|E_i^*|\) must lie between these two values.

**Corollary 1.1.** At the optimum, group \(i\) will acquire (i) all the learnable elements from its neighbors with degree smaller than or equal to \(d^*\); (ii) a proper subset of learnable elements from its degree \(d^*+1\) neighbors; and (iii) no elements from its neighbors with degree greater than \(d^*+1\).

**Proof.** This follows directly from Lemma 1.1 and Proposition 1.2.

Furthermore, if the region’s dialects satisfy a symmetry condition, then Proposition 1.2 immediately implies that all groups learn the same measure of elements in a typical period.

**Definition 1.4** (Symmetry). The set of dialects \(\{L^t_i\}_{i \in \mathcal{G}}\) is symmetric at \(t\) if the following two conditions are satisfied:

\[
|L^t_i| = |L^t_j|, \quad \forall i, j \in \mathcal{G} \quad (1.15)
\]

\[
|L^t_i \cap L^t_j| = |L^t_i \cap L^t_k|, \quad \forall i, j, k \in \mathcal{G} \text{ such that } d_{ij} = d_{ik}. \quad (1.16)
\]

Condition (1.15) requires that the measure of each dialect is the same. Condition (1.16) requires that the intersection of any dialect \(L^t_i\) with that of any equally distant neighbors has the same measure.

**Corollary 1.2.** If the set of dialects \(\{L^t_i\}_{i} \) is symmetric and localized, then the optimal size of \(E^*_i\) is the same for all groups in any given period, i.e. \(|E^*_i| = |E^*|\) for all \(i \in \mathcal{G}\).

**Proof.** This follows directly from the properties of symmetry and localization of the set of dialects, and the symmetry of each group’s decision.

While Proposition 1.2 makes no requirement about which subset of the degree \(d^*+1\) neighbors’ dialects a group should learn, we make two assumptions about how groups learn from their \(d^*+1\) neighbors’ dialects. The first assumption ensures that the set of dialects stay symmetric and localized in the subsequent period, so that results derived in this section can be applied to analyze long-run properties of the region’s languages in a dynamic context. The second assumption enables us to use the definition of same language (i.e. Definition 1.2) consistently in the dynamic setting.

**Assumption 1.1.** When a group is indifferent between learning from several dialects, it will learn the same measure of elements from each of them; the elements learned are chosen randomly from the set of learnable elements in those dialects.

\[37\] Without Assumption 1.2, we would have to modify the criterion for same language to a slightly more complicated version: \(L^t_i \cup \left( \bigcup_{k=1}^{d^*} E_{ik} \right) = L^t_j \cup \left( \bigcup_{k=1}^{d^*} E_{jk'} \right)\). The other aspects of the model are unaffected by this assumption.
**Assumption 1.2.** All groups within a degree $d^*$ neighborhood learn the same subset of elements from their degree $d^* + 1$ neighbors.

Notice that the characteristics of the set of learnable elements may vary from period to period. Consequently, unlike $|N^*|$, $|E^*|$ need not be the same across time.

### 1.5 Long Run

In this section, we examine the long run implications of the model. Suppose that a period represents a human generation of approximately twenty years. Each new generation of group members inherits the dialect of the previous generation, and then may choose to learn from the neighbors’ dialects as well as to invent novel expressions.

We assume that groups are myopic, so that each generation is only interested in payoffs in the current period. While this assumption frees us from considering repeated game effects, the groups’ myopia over payoffs is not implausible. Group membership in small scale societies is fluid across generations. Most groups are patrilocal or matrilocal, so adults expect that approximately half of their offspring, at maturity, will leave for another group. This consideration reduces the incentives of the current generation to consider the costs and benefits of their choices on future generations of the group. Myopic play implies that each group’s decision in any given period is characterized in Section 1.4. This allows us to analyze the trajectory of a region’s linguistic composition over generations.

**Definition 1.5.** The linguistic composition in the region at time $t$ is the set

$$
\Lambda^t = \left\{ (\lambda, \mathcal{G}_\lambda) : \lambda = \bigcup_{k \in \mathcal{G}_\lambda} (L^t_k \cup E^t_k \cup N^t_k) \text{ where } \mathcal{G}_\lambda \text{ is } \right\}
$$

the set of groups speaking the same language $\lambda$.

Recall that the criterion for “same language” is given in **Definition 1.2.** We are interested in the steady state of a region’s linguistic make-up; that is, a linguistic composition that is stable in the long run. We denote such a composition $\overline{\Lambda}$, wherein each language $\lambda \in \overline{\Lambda}$ is spoken by the same subset of groups $\mathcal{G}_\lambda \subseteq \mathcal{G}$ in every subsequent period.\(^{38}\) Henceforth, we use a superscript $t$ on the variables to index time. In accordance with **Definition 1.2,** we formally define a steady state as follows:

**Definition 1.6.** The region’s linguistic composition is in a steady state if for all $t$,

$$
L^t_i \cup E^t_i = L^t_j \cup E^t_j \implies L^{t+1}_i \cup E^{t+1}_i = L^{t+1}_j \cup E^{t+1}_j
$$

\(^{38}\)Note however that the languages themselves will not be the same; they will grow in size over time, according to the results in Section 1.4.
and
\[ L_i^t \cup E_i^t \neq L_j^t \cup E_j^t \quad \Rightarrow \quad L_i^{t+1} \cup E_i^{t+1} \neq L_j^{t+1} \cup E_j^{t+1} \] (1.18)
for all \( i, j \in \mathcal{G} \).\(^{39}\)

A linguistic composition is in a steady state if the following two conditions hold: (i) every pair of groups \( i, j \) that speak the same language in any period \( t \) will continue to speak the same language in \( t + 1 \); and (ii) every pair \( i, j \) that do not speak the same language in \( t \) will not speak the same language in \( t + 1 \). Therefore, a steady state of the linguistic composition is a set \( \Lambda \) of languages, each of which is spoken by the same subset \( \mathcal{G}_\lambda \) of groups over time.

Recall from Section 1.3.3 that the order of events in a typical period \( t \) is as follows:

i) \( \{|L_j^t \setminus L_i^t| : i, j \in \mathcal{G}\} \) is observed by all groups.

ii) \( \{E_i^t\}_i \) are chosen simultaneously.

- The number of languages for period \( t \), \( \#(\Lambda^t) \), is counted at this point.\(^{40}\)

iii) \( \{N_i^t\}_i \) are chosen simultaneously.

- The set of dialects at the beginning of period \( t + 1 \), \( \{L_i^{t+1}\}_i = \{L_i^t \cup E_i^t \cup N_i^t\}_i \), is determined at this point.

iv) A period game is drawn for every possible pair of groups according to (1.5) and (1.8).

v) The period games are played, and payoffs are determined based on \( \{L_i^t, E_i^t, N_i^t\}_i \).

For the results in Section 1.4 to apply in every period, it must be the case that the set of dialects is symmetric and localized at the beginning of every period. The following lemma shows that if \( \{L_i^t\}_i \) satisfies symmetry and localization in the initial period, it will continue to do so in subsequent periods.

\textbf{Lemma 1.2.} Suppose \( \{L_i^1\}_i \) is symmetric and localized. Then \( \{L_i^{t+1}\}_i \) is also symmetric and localized.

The proof of this lemma can be found in Section A.2.1.

Before we state our main dynamic results, it helps to define an important value, \( \overline{d} \). As we will state formally in Proposition 1.4 and Proposition 1.5, \( \overline{d} \) is the threshold value of neighborhood distance. Groups with \( d_{ij} \leq \overline{d} \) will speak the same language in the steady state, and groups with \( d_{ij} > \overline{d} \) will speak different languages.

\textbf{Definition 1.7.} Define \( \overline{d} \) as an element of \( \{0, \ldots, D\} \) that satisfies the following condition:

\( ^{39} \)In our model, the sizes of the languages are growing over time. Hence, the steady state of a linguistic composition is steady in the sense that each language in the composition has a stable speakership.

\( ^{40} \)\( \#(\Lambda) \) is the number of elements in \( \Lambda \).
\begin{equation}
\frac{r + \beta}{1 + r} \pi(d) \geq \sum_{k=0}^{d} |P^t_i(k)| \text{ and } \frac{r + \beta}{1 + r} \pi(d + 1) < \sum_{k=0}^{d+1} |P^t_i(k)|,
\end{equation}

where
\[|P^t_i(k)| = \begin{cases} 
2^{k-1}|N^*| & \text{if } k \in \{1, \ldots, d\} \\
0 & \text{if } k = 0
\end{cases}\]
and \(|P^t_i(k)| \geq 2^{k-1}|N^*| \text{ if } k \geq d + 1.\]

The particular neighborhood distance \(d\) has some special properties. The next proposition, for example, establishes that if \(d_{ij} > d\) for a pair of groups \(i, j\), then their languages will, in some sense, become less similar over time. This is because the size of the set that \(j\) invents every period, \(|N_j|\), will always be larger than the size of the set that \(i\) learns from \(j\), \(|E_{ij}^*|\). With each passing period, the set of elements that \(i\) speaks but \(j\) doesn’t will become larger. Due to symmetry, the set of elements that \(j\) speaks but \(i\) doesn’t will also become larger.

**Proposition 1.3.** If \(d_{ij} > d\) then \(|E_{ij}^*| < |N_j|\). Thus, \(|L_j \setminus L_i|\) increases every period.

The proof can be found in Section A.2.2. Proposition 1.3 is a result regarding the trajectory of a pair of languages. From (1.5), \(\pi(d)\) is the probability of a non-null interaction occurring between the groups. According to (1.6), \(\pi(\cdot)\) is a decreasing function of \(d\), so the further apart \(i\) and \(j\) are geographically, the less they interact. Proposition 1.3 establishes that if meaningful interactions between \(i, j\) are sufficiently infrequent, that is, if \(d\) is high enough, their dialects will become less and less similar over time. This process might look empirically similar to linguistic drift, as both types of divergence theoretically decrease with frequency of interaction.

The next proposition builds on the results of Proposition 1.3, establishing that if \(d_{ij} \leq d\), and \(i, j\) begin by speaking the same language, then their languages will continue to be the same. That is, in every period \(i\) will learn all of \(j\’s\ newly invented elements, and vice versa.

**Proposition 1.4 (Existence and characterization of a steady state).** There exists a steady state, \(\Lambda\), which is characterized by the following condition:
\begin{equation}
L^t_i \cup E^t_i = L^t_j \cup E^t_j \iff d_{ij} \leq d, \quad \forall i, j \in G,
\end{equation}
where \(d \in \{0, \ldots, D\}\) is determined by (1.19).

The proof of this proposition can be found in Section A.2.3. The proposition establishes that there is a steady state of the region’s linguistic composition which can be fully described by a particular degree of neighborhood, \(d\). In this steady state, all groups that are members of the same \(d\) neighborhood speak the same language, and furthermore only groups which
are members of the same \( \bar{d} \) neighborhood speak the same language. In other words, if \( d_{ij} \leq \bar{d} \) then \( L_i \cup E_i = L_j \cup E_j \), and if \( d_{ij} > \bar{d} \) then \( L_i \cup E_i \neq L_j \cup E_j \). Note that \( \bar{d} \) depends only the exogenous parameters of the model (\( |N^*| \) is also a function of those parameters). Thus, \( \bar{d} \) is constant over time.

**Corollary 1.3** (Steady state number of languages). *The number of languages in the steady state is \( \#(\Lambda) = G/2^{\bar{d}} \).*

*Proof.* This follows directly from condition (1.20). Each degree \( \bar{d} \) neighborhood is inhabited by \( 2^{\bar{d}} \) groups, and these groups share a common language in the steady state. Since there are \( G \) groups, the number of languages in the steady state is therefore \( G/2^{\bar{d}} \).

Next we turn our attention to convergence, showing that for a general set of initial conditions, the linguistic composition converges towards the \( \Lambda \) described in Proposition 1.4.

**Proposition 1.5.** If \( \{L_i^0\}_{i \in G} \) is symmetric and localized, then after a finite number of periods, the steady state \( \Lambda \) is reached.

The proof of this proposition is presented in Section A.2.4. This proposition says that, after finitely many periods, \( i \) will learn the entire set of learnable elements from its degree \( \bar{d} \) neighbors and will therefore, by definition, speak the same language as each of them. It is worth noting the two special initial linguistic compositions where all groups have identical initial dialects or completely distinct ones both satisfy the symmetry and localization conditions.

Every period, \( i \) learns a larger set of the languages of its \( d \leq \bar{d} \) neighbors than those groups collectively invent. In other words, \( |E_i^*| > 2^{\bar{d} - 1}|N^*| \). Suppose \( i \) does not initially speak the same language as its \( \bar{d} = 1 \) neighbor, and \( \bar{d} \geq 1 \). In the first period, \( i \) begins “catching up” with its \( \bar{d} = 1 \) neighbor, then later with its \( \bar{d} = 2 \) neighbors, and so forth until it has learned the entire set of learnable elements from its neighbors of degree \( d \leq \bar{d} \). At this point, \( L_i \cup E_i = L_j \cup E_j \) holds for all of \( i \)'s \( d \leq \bar{d} \) neighbors.

From this point onwards, \( i \) learns at least \( 2^{\bar{d} - 1}|N^*| \) elements every period, including all those invented in the last period by its neighbors of degree \( d \leq \bar{d} \). \( i \) may learn some elements from its degree \( \bar{d} + 1 \) neighbors, but as was shown in Proposition 1.3, not as many as this set of neighbors invents. For this reason, \( i \) never catches up with these neighbors.\(^{41}\)

From Corollary 1.3, \( \#(\Lambda) = G/2^{\bar{d}} \), therefore \( G/2^{\bar{d}} \) is the steady state number of languages. Last but not least, since the steady state is uniquely determined by \( \bar{d} \), we conclude that it is the unique steady state for the set of initial linguistic compositions that we consider.

\(^{41}\)The divergence of \( i \)'s language from those of its higher degree neighbors actually occurs immediately, as a result of its choice of \( E_i \) in the first period.
1.6 Comparative Statics

Now we return to our original purpose: to show the effect of cooperative and competitive incentives on the number of languages in a region. In this section, we establish the key comparative static result for the number of languages in the steady state: \( \#(\overline{\Lambda}) \) is weakly decreasing in \( r \), the ratio of the probabilities of cooperative and competitive interactions.

Furthermore, we examine the results of exogenous changes in the geographic locations of groups on the relationships between particular dialects. Specifically, we determine under what circumstances such a change would alter the steady state identity of the dialects of an arbitrary pair of groups.

**Proposition 1.6.** The steady state number of languages, \( \#(\overline{\Lambda}) \), is weakly decreasing in \( r \).

In other words, the higher the regional ratio of the probability of cooperation vs. competition, the fewer languages there will be in the region in steady state.

**Proof.** According to Corollary 1.3, \( \#(\overline{\Lambda}) = G/2^d \). We show that \( \#(\overline{\Lambda}) \) is weakly decreasing in \( r \) by proving that \( d \) is weakly increasing in \( r \). Recall that \( d \) is determined by (1.19). In particular, it must be the case that, for any \( d \in \{1, \ldots, D\} \), the first inequality of (1.19) can be rearranged as follows:

\[
\frac{r + \beta}{1 + r} \pi(d) \geq (2^d - 1)N^* \quad \text{by (1.9)}
\]

Similarly, for any \( d \in \{0, \ldots, D - 1\} \), the second inequality of (1.19) can be written as

\[
(r + \beta) < \frac{2^{d+1} - 1}{\pi(d+1)} \left[ \beta \sum_{k=1}^{D} 2^{k-1} \pi(k) \right]. \quad (1.22)
\]

(1.21) and (1.22) together imply that

\[
\frac{2^d - 1}{\pi(d)} \left[ \beta \sum_{k=1}^{D} 2^{k-1} \pi(k) \right] \leq r + \beta < \frac{2^{d+1} - 1}{\pi(d+1)} \left[ \beta \sum_{k=1}^{D} 2^{k-1} \pi(k) \right], \quad (1.23)
\]

where the terms in the square brackets are independent of either \( r \) or \( d \). Note that the fraction \( (2^k - 1)/\pi(k) \) is strictly increasing in \( k \), because \( \pi(\cdot) \) is decreasing. Thus, when the increase in \( r \) is sufficiently large, \( d \) would also have to increase in order for (1.23) to hold. Likewise, when the decrease in \( r \) is large enough, \( d \) needs to decrease as well to satisfy (1.23). This completes the proof. \( \Box \)
Proposition 1.7. The steady state number of languages, $\#(\Lambda)$, is not affected by multiplying the probability of a non-null interaction, $\pi(d)$, by any positive constant, $\phi > 0$.

Proof. The statement follows trivially from (1.23): note that a constant $\phi$ multiplying $\pi(\cdot)$ gets canceled out immediately.

This result shows that an exogenous change in the frequency of meaningful interactions in the region, as long as it does not affect either the relative frequency of cooperative vs. competitive interactions, $r$, nor the set of ratios $\pi(d)/\pi(d+1)$, $d \in \{1, \ldots, D-1\}$, will have no effect on the steady state number of languages in the region.

1.7 Discussion

As a means of transmitting information between human beings, nothing can be more fundamental than language. Yet languages themselves are complex and dynamic. Understanding the mechanism behind linguistic change, therefore, is an interesting and worthwhile undertaking. In this paper, we propose that the convergence and differentiation of languages can be partly explained in terms of variations in the strategic nature of the environment in which linguistic groups reside. In our model, languages homogenize when cooperation prevails in a region, and that languages become more numerous as more conflicts arise.

Most noteworthy in our model is that divergence of languages is endogenously generated. This explanation contrasts with the conventional explanation of linguistic differentiation as a result of isolation and drift. Both theories predict that the languages of a pair of geographically distant groups will become less similar over time. The distinction is that with drift, the change is due to random mutation, whereas in our model, each group chooses to differentiate their language from all neighbors. Since $d_{ij} > \bar{d}$, $i$ and $j$ do not interact frequently enough to make it worthwhile to keep up with the changes in each other’s language via learning.

The context of our model—small scale farming or foraging economies—may cause some readers to overlook its relevance to contemporary linguistic diversity. It is important, however, to note that linguistic change is usually slow (Pagel et al., 2007). Sometimes a generation may not even be conscious about the changes occurring in their language. Anecdotal evidence of the English language suggests that it was already close to its present form before the advent of modern transportation and communication technology. An average English speaker today, for example, has little trouble understanding Jonathan Swift’s *Gulliver’s Travels*, written in 1726, decades before the industrial revolution. This time lag in the determination of linguistic diversity suggests that we should not expect much contemporaneous effect of current prevalence of trade and/or conflict on current linguistic diversity. Despite our focus on small-scale societies, we hypothesize that cooperative and competitive incentives are important causes of linguistic change in the modern world. The
factors affecting the type and frequency of interaction, however, are complex and change quickly.
Chapter 2

Strategic Incentives and Language Change

2.1 Introduction

In Chapter 1, Chen and Mitchell (henceforth CM) propose a theory that delineates how linguistic diversity, i.e. the number of languages in a region, may be partly determined by the relative frequency of the cooperative and competitive interactions in which autonomous groups in that region engage. Specifically, CM’s theory predicts that a high frequency of cooperative interactions—modeled as linguistic groups interacting in coordination games—would lead to homogeneity in the groups’ languages, hence resulting in low linguistic diversity; and a high frequency of competitive interactions—modeled as zero-sum games—would cause the groups’ languages to diverge, thus resulting in high linguistic diversity. To test CM’s predictions using conventional econometric methods would require data on linguistic diversity observed over multiple time periods. Unfortunately, such a dataset is rather difficult to compile. As an alternative solution, I propose in this paper an economic experiment to test a main hypothesis from CM’s paper: cooperation leads to linguistic homogeneity, and competition leads to linguistic divergence.

2.2 The Chen-Mitchell Model

In CM’s model, $G$ groups live in a region with neighborhoods structured like a binary tree (see Figure 2.1 for an illustration). The shortest traveling distance for groups $i$ and $j$ to meet in a common middle location is denoted by $d_{ij}$, with $d_{ij} \in \{0, 1, \ldots, D\}$ where $D$ is the farthest such distance for two groups in the region.\footnote{Thus, $G = 2^D$ and $D = \log_2 G$. Also, $d_{ij} = 0$ if and only if $i = j$.} A neighborhood of degree $d$ consists of a set $\mathcal{G}$ of groups such that for every $i, j \in \mathcal{G}, d_{ij} \leq d$. As is shown in Figure 2.1, a degree $d$ neighborhood has $2^d$ groups.
All groups interact periodically in pairs, in one of three possible games: coordination, zero-sum, and null. These games are used to model cooperative, competitive and no interactions, respectively. The probability of groups $i$ and $j$ playing a coordination game is $p_{ij}$, the probability of them playing a zero-sum game is $q_{ij}$, and the probability of them playing a null game is $1 - p_{ij} - q_{ij}$. The region’s geographic conditions, summarized by a constant parameter $r$, determine the relative probability of coordination versus zero-sum games: $r = p_{ij}/q_{ij}$ for any $i, j$ pair. The farther apart two groups are, the more likely they are to interact in a null game. CM capture this by assuming $1 - p_{ij} - q_{ij} = 1 - \pi(d_{ij})$, where $\pi(d)$, the probability of two degree $d$ neighbors having any interaction at all, is decreasing in $d$ and at a sufficiently fast rate.

Each group speaks a dialect—a set of words. Two dialects are considered to be variants of the same language if they share a sufficient amount of common words; that is, if the number of elements in the intersection of the two dialects is large enough.

The timing of events within a typical period is as follows.

i) Each group observes how much it does not know about the other groups’ dialects.

ii) Each group makes a decision about how much of the other groups’ dialect to learn, and how many new words to invent.

iii) Groups are then randomly matched to play one of the three games described above.

---

2A null game is one where the payoffs are always zero regardless of the actions taken.

3Specifically, CM assume $\pi(d) \geq 2^d \pi(d + 1)$.

4Groups in CM’s model make decisions to change their dialects in every period: either by learning words from existing dialects or by inventing new words. Let $L_k^t$ denote group $k$’s dialect at the beginning of $t$; let $E_k^t$ be the set of words learned by $k$ at $t$ from other existing dialects, and $N_k^t$ the set of words $k$ invents at $t$; let $L_k^{t+1} = L_k^t \cup E_k^t \cup N_k^t$ be $k$’s dialect at the beginning of $t + 1$. Then, $L_i^{t+1}$ and $L_j^{t+1}$ are considered variants of the same language if $L_i^t \cup E_i^t = L_j^t \cup E_j^t$. See Section 3.2 of CM’s paper for a more detailed explanation.
iv) Games are played, and payoffs are realized.

This period game is played repeatedly by the groups, which are assumed to be myopic. A group’s expected payoff in a period game is a function of the relationship between its own dialect and its opponent’s dialect.

In a coordination game, group $i$’s payoff is increasing in the commonality between $i$ and $j$’s dialects. This captures the intuition that coordination is more likely to be successful when the two groups can communicate in an accurate and inexpensive manner. During a conflict, on the other hand, $i$’s expected payoff increases in how much it understands the opponent $j$’s dialect, while it decreases in the level of $j$’s comprehension of $i$’s dialect. Intuitively, when $i$ and $j$ are in an antagonistic relationship, the more information $i$ has about $j$, the more advantaged $i$ would be in that relationship. Since language is the most common medium of intra-group communication, a better understanding of $j$’s dialect leads to a higher chance of obtaining $j$’s private information. By the same token, the less $j$ understands $i$’s dialect, the better $i$ would be able to protect its own secrets.\footnote{When concealing sensitive information, differentiated language has an additional benefit: it helps group members identify each other.} Lastly, if two groups do not interact in a period, which is modeled as a null game, then their payoffs are unaffected by the relationship between their dialects.

The functional forms of these expected payoffs are determined by an underlying communication game: depending on the type of interaction, the groups play a symmetrized version of a coordination, zero-sum, or null game with one-sided pre-play communication. CM offer the following interpretation to this structure: Suppose each group has a leader who is in charge of coordinating the actions of the group members. Before each interaction with another group, the leader gives an instruction to her group members about what to do in the game. The group members would then carry out the leader’s instruction. The one-sided pre-play communication can be thought of as the instruction of one group being overheard by the other group.

Each group’s dialect is modeled as the message space in the underlying communication game. Group $i$ will understand group $j$’s message only if that message is also an element in $i$’s dialect. After some normalization of the payoffs, when the underlying game $\Gamma$ is a coordination game, a group’s expected payoff in the Pareto efficient Nash equilibrium increases with the number of common elements in their dialects; and when $\Gamma$ is a zero-sum game, group $i$’s expected payoff in the Nash equilibrium is increasing in the number of elements that $i$ knows but $j$ doesn’t and decreasing in the number of elements that $j$ knows but $i$ doesn’t.

CM show that as the period game is played for a sufficiently long period of time, there exists a number $\bar{d}$, as a function of the parameter $r$, such that all groups within a degree $\bar{d}$ neighborhood will share the same language.
Proposition 2.1. There exists $d \in \{0, \ldots, D\}$ such that, for $t$ sufficiently large,

$$L_i^t \cup E_i^t = L_j^t \cup E_j^t \iff d_{ij} \leq d.$$  \hspace{1cm} (2.1)

In other words, if two groups, $i$ and $j$, are within a degree $d$ neighborhood, the amount of words learned by $i$ from $j$ is no smaller than the amount of words $j$ invents within a period. By implication, the dialects of groups living outside a degree $d$ neighborhood will become less and less similar with each passing period.

Corollary 2.1. If $d_{ij} > d$, the amount of words learned by $i$ from $j$ ($|E_{ij}^t|$) is less than the amount of words invented by $j$ ($|N_j^t|$), i.e. $|E_{ij}^t| < |N_j^t|$ for $t$ sufficiently large.

As a result, there will be $G/2^d$ languages in the region. CM call this the steady state number of languages. The most relevant result for the present paper is the following:

Proposition 2.2. As $r$ increases, namely, as cooperation becomes more likely relative to competition, the steady state number of languages in the region weakly decreases.

The next section proposes an experimental design to test this proposition.

2.3 Experimental Design

This section describes the experimental design to test the theoretical predictions of the model. Notice that Proposition 2.2 is essentially a comparative statics statement. Therefore to test it, I utilize a $2 \times 2$ design. The primary treatment variable controls the nature of interactions: in a COOPeration treatment, groups only interact in coordination games, and in a COMPetition treatment, groups only interact in zero-sum games. The secondary treatment variable is the initial linguistic similarity of the interacting groups: they may start out with identical dialects or distinct ones. This is used to control for possible effects of initial conditions.

To keep things simple, I focus on the case of two-group interactions ($G = 2$) and test the trend in the changes of the groups’ dialects over time. As I am especially interested in the strategic incentives, rather than the lack thereof, on linguistic change, I set the probability of null interaction to zero. Hence, $d$ is either 0 or 1. Proposition 2.2 says that $d$ weakly increases as $r = p_{ij}/q_{ij}$ increases. Thus, in the COOP treatment (where $p_{ij} = 1$ and $q_{ij} = 0$), $d = 1$, and in the COMP treatment (where $p_{ij} = 0$ and $q_{ij} = 1$), $d = 0$. Additionally, Corollary 2.1 implies that groups learn less from neighbors living outside a degree $d$ neighborhood than the latter would create every period. Hence, I derive and test the following hypothesis.

Hypothesis 2.1. In the COOP treatment, the amount of words learned exceeds the amount of words created, i.e. $|E_{ij}| - |N_j| \geq 0$. In the COMP treatment, the amount of words learned is less than the amount of words created, i.e. $|E_{ij}| - |N_j| < 0$.

\footnote{The steady state is “steady” in the sense that each language has a stable speakership.}
2.3.1 Subjects and Treatments

Subjects of the experiment are recruited using the SFU Centre for Research in Adaptive Behaviour in Economics (CRABE) online recruiter. In CM’s paper, the basic decision unit is a group, which has a leader who issues instructions and group members who carry out those instructions. Since the interests of the leader and the group members are perfectly aligned, there is no strategic reason for a member to not obey the leader’s command. Therefore, to make more efficient use of the subjects, instead of having more than two human subjects forming a group, I let each subject assume the role of a group’s leader and let the group members be a computerized agent that faithfully executes the subject’s choices. Letting group members be computerized agents also minimizes the potential noise in the data added by human subjects due to misunderstanding or other non-salient factors.

In this experiment, I choose \( G = 2 \), so that each “region” only has two groups (i.e. two subjects). Each session has an even number of subjects (with a minimum of four subjects), who are assigned to the roles of either Person RED or Person BLUE. A Person RED is then matched to a fixed Person BLUE throughout the experiment, although the subjects are not told about this arrangement. This design has two objectives. First, as will be clear in Section 2.3.5, I rotate pairs of dialects across pairs of subjects so that the dialects may evolve as if they were modified and used by generations of myopic agents. Such a rotation, when combined with the random re-matching of subjects, however, would create the possibility that a subject is endowed with the same dialect in two consecutive periods. This implies that subjects may be able to learn words without having to pay for them within a period. Fixed matching is a solution to this problem. Second, not telling the subjects that they are matched with the same counterpart throughout helps alleviate the concern for subjects adopting history-dependent strategies, which are plausible if subjects know that they will be interacting repeatedly with the same counterpart.

There are two treatments in this experiment. In the COOP treatment, subjects always play a coordination game, while in the COMP treatment, subjects always play a zero-sum game. I set the probability of a null interaction to zero, so as to focus on the contrast between the effects of cooperative and competitive interactions on language change. Further, there are two variations in each treatment: each pair of subjects start out either with exactly the same dialect or with completely different dialects.

2.3.2 Timeline of a Session

Figure 2.2 illustrates how a typical experimental session proceeds. After the subjects are seated, the experimenter reads the instructions. Next, the subjects are given two practice

---

7 In the experiment instructions (see Section B.1), subjects are told that they will be participating in a two-person interaction. Thus, the narrative of the subjects being group leaders is only for the purpose of interpretation.
rounds to help them get familiar with the experimental procedure. Then, the subjects complete a quiz composed of ten true-or-false questions, aimed at helping them better understand the procedure. The answer and its explanation to each question is displayed immediately after a subject submits her response, so that the subject could know not only the correct answer but also why that answer is correct. Moreover, the subjects get paid for each correctly answered quiz question, and they are told this at the beginning of the instructions. According to Freeman et al. (2017), paid quizzes and providing explanations to answers can significantly enhance subjects’ understanding of the experimental procedures.

The main part of a session consists of 20 rounds of a “period game” resembling the one in CM’s model. The subjects are not informed of the total number of rounds.

A typical round has two phases. In Phase 1 of each round, subjects decide how to modify their “dialects”, called message lists, which contain messages that refer to actions in the interaction in Phase 2. The message lists of a matched pair of subjects have a common portion called public list, akin to the intersection of two dialects in CM’s model, and messages not in the public list belong to each Person’s private list. In Phase 1, a subject can either learn messages (moving messages in the other Person’s private list to the public list) or create new messages (adding new messages to her own private list). Then in Phase 2, a matched pair play either a $3 \times 3$ coordination game or a $3 \times 3$ zero-sum game, prefixed by a one-sided pre-play communication. Specifically, each subject first decides which action in the $3 \times 3$ game to take by submitting a corresponding message to the computer; upon receiving the two messages, the computer randomly reveals one Person’s message to the other Person who would then have the chance to revise her initial choice of action based on her counterpart’s submitted message. After the revision is completed, a typical round ends. Before a new round begins, a subject can observe the outcome of the Phase 2 game that she just played and the calculation of her payoff of the current round.

2.3.3 Dialects and Linguistic Similarity

Dialects in CM’s model are operationalized as lists of messages that refer to actions in the Phase 2 game, as shown in Figure 2.5. To avoid priming effects, I use symbols from the
Figure 2.3: Sample Message List for a Subject

Unicode character table that have no apparent relations with the letters “A” “B” or “C”, the names of the three actions in the Phase 2 game. Each message (i.e. symbol) refers to one and only one of the three actions.\(^8\) If we think of a message list as a dictionary, then the Unicode symbols would be words and the meanings of those words are either A, B, or C.

At the beginning of every round, each subject is endowed with a possibly different set of messages, divided into two lists, as shown in Figure 2.3. The public list is the same for both subjects in a matched pair: it contains the same Unicode symbols and the meanings of those symbols are the same for both subjects. The private list is unique to each subject: the meanings of the symbols therein are known only by that subject. In a given list, either public or private, it is possible to have no message referring to a particular action (e.g. the private list in Figure 2.3 has no message referring to actions A or B); it is also possible to have multiple messages referring to an action (e.g. the public list in Figure 2.3 has two messages referring to action A). However, when a subject’s public and private lists are taken together, she always has at least one message that refers to each of the three actions at the beginning of the first round.\(^9\)

\(^8\)In this design, messages are only allowed to refer to single actions for two reasons. First, it makes the experiment easier to be explained to the subjects. Second, a message that refers to more than one action is essentially the same as asking the computer to randomly pick one of the actions according to some probability distribution. However, it is the subject (i.e. the group leader), not the computer (the group member), who should make decisions about such a randomization. The computer should do nothing more than executing a decision made by the subject. Therefore, if the need for randomization arises, it should be resolved with the subject, not left to the computer. It may nevertheless appear reasonable to have a “null” message that means “choose whichever action you want” or “I’m not going to tell you what to do”. Such a null word is essentially a word that refers to the entire set of actions. But note that it would be against a subject’s best interest to deliberately obfuscate the meaning of a message sent to the computer, since the leader and her group members have perfectly aligned interests. If, out of strategic consideration, a subject \(i\) wants the other subject \(j\) she is paired with to see this null message (so that \(j\) couldn’t predict \(i\)’s decision with certainty even if \(j\) intercepts that message), then \(i\) could have achieved the same effect by using a message in the private list.

\(^9\)The subjects are not informed of this design feature. Nevertheless, suppose a subject somehow believes correctly that one has to have at least one message for every action. Then it is possible in principle for him to infer the meaning of his counterpart’s private message when it is the only message in the latter’s private list and there is no corresponding public message that has the same meaning. From the experimental data, however, I do not observe message lists that match this pattern.
In Phase 1 of each round, a subject can learn and create messages through a user interface (UI) similar to Figure 2.4. There are four selection boxes in the bottom half of the screen. Learning is done by choosing a number in the first selection box. A subject can learn as many messages as there are in the other person’s private list. The number of messages in the other person’s private list is shown on the screen as “(AVAILABLE: #)”. Creating messages is done by choosing numbers in the bottom three selection boxes, each corresponds to a particular action. A subject is allowed to create at most three new messages in each round. Both learning and creating messages are costly. Specifically, the cost of either activity is given by $\frac{1}{2}x(x + 1)$, where $x$ is the number of messages learned or created. Thus, if a subject learns two messages and creates three, the total cost incurred in Phase 1 would be $3 + 6 = 9$. The message lists will update after the subjects submit both the decisions to learn and to create. Therefore, a message created in a given round cannot be learned in that same round.

2.3.4 Period Game (Phase 2 Interaction)

The Phase 2 interaction is essentially a symmetrized version of the games in Figure 2.5 with one round of pre-play communication. Depending on the treatment, subjects either play

---

10 This upper limit on message creation is imposed because of a shortage of Unicode symbols that can be properly displayed by the computer, and because there is no apparent benefit to creating more than three messages given the convex costs.
a coordination game similar to the one in Figure 2.5(a) or the zero-sum game featured in Figure 2.5(b). Each subject in a matched pair first submits an intended action through a screen similar to Figure 2.6(a) by selecting a message associated with that action. After receiving submissions from both subjects, the computer (uniformly) randomly selects one subject’s submission and reveals it to the other subject, as shown in Figure 2.6(b). The latter can then revise her initial choice based on the other subject’s message. If the former subject’s message is from the public list, the subject who gets the revision opportunity will see action chosen by the other in the “ACTION INTENDED” field; otherwise, the “ACTION INTENDED” field will display “UNKNOWN” as in Figure 2.6(b). The outcome of the interaction is determined by the second choice of the subject who gets the revision opportunity and the initial choice of the other subject.

2.3.5 Capturing the Evolution of Languages

In CM’s model, generations of myopic agents make changes to their dialects through learning and creating words. To capture this dynamic aspect of the model, I implemented the following design that rotates pairs of message lists between rounds across pairs of subjects.

Each session is run with an even number of (and at least four) subjects, say, $S_1, \ldots, S_{2n}$, who are matched in pairs. There are also $2n$ message lists, separated into $n$ pairs. Thus, in any particular round $R$, $n$ pairs of subjects are making changes to $n$ pairs of message lists independently. Then in round $R+1$, I rotate the pairs of message lists between different pairs of subjects so that each pair of subjects never modify the same set of message lists in two consecutive rounds. For example, suppose in round $R$, the pair of subjects $S_1$ and $S_{n+1}$ are given message lists $V_1$ and $V_{n+1}$, $S_2$ and $S_{n+2}$ are given $V_2$ and $V_{n+2}$, and likewise for other subjects. Then in $R+1$, $S_1$ and $S_{n+1}$ would be given $V_2$ and $V_{n+2}$, $S_2$ and $S_{n+2}$ are given $V_3$ and $V_{n+3}$, and similarly for the rest of the subjects.

Figure 2.7 illustrates how message-list rotation works in a session with eight subjects. Suppose in a particular round $R$, we see that subjects $S_1, \ldots, S_8$ are using message lists

---

![Figure 2.5: Phase 2 Games](image)

(a) Phase 2 Game in COOP

(b) Phase 2 Game in COMP

---

11 The locations of the payoff profiles vary across rounds, although the payoffs themselves remain unchanged. Moreover, in all the variant games, the two pure strategy equilibrium payoff profiles are attained by distinct action profiles. That is, if the pure action profiles $(a_i, a_j)$ and $(a'_i, a'_j)$ both attain the $(40, 40)$ payoff, then $a_k \neq a'_k$ for $k = i, j$. 

---

39
(a) Phase 2 Decision—Initial Choice

(b) Phase 2 Decision—Revision Opportunity

Figure 2.6: Phase 2 Decisions
L₁, . . . , L₈, respectively. Then, in round R + 1, the rotation mechanism would assign L₁ (previously used by S₁) to S₃, L₂ (previously used by S₂) to S₄, and so on. This design allows me to track the evolution of each pair of vocabularies over time, as if they were modified and used by different generations of myopic agents. In the instructions, subjects are told that “[t]he set of messages given to you will be different in every round.”

The rotation mechanism ensures that subjects cannot use the same message lists in any two consecutive rounds. However, it is still possible that, after enough rounds of rotations, they are assigned the same message lists that they used a few rounds ago. Consequently, subjects may be able to learn the meaning of a word without paying the associated cost. This issue would be especially pronounced in sessions that have only four subjects, where such a turn-around time is every two periods. To address this problem, all symbols in the message lists would be replaced by a new set of symbols during a rotation, while the configuration of the vocabularies is preserved. To illustrate, if before the end of a round R, a vocabulary looks like

then, at the beginning of round R + 1, it will look like
The number of symbols in each list and under each heading remain the same, but the original symbols are replaced by new ones after the rotation. This measure takes away a subject’s dynamic incentives in making decisions.

### 2.3.6 Payment

At the end of a session, a subject receives payment based on the following three items:

- $7 for showing up on time,
- the final payoff (in Experimental Dollars, or ED) of one randomly selected round,
- 1 ED per correctly answered quiz question.

Each subject starts a round with an initial payoff of either 15 ED (in COOP) or 30 ED (in COMP). The final payoff of a round is determined by the following formula:

\[
\text{initial payoff} - \text{costs of learning and creating words} + \text{interaction payoff.} \quad (2.2)
\]

The calculation of the round payoff is shown to the subjects at the end of each round (see Figure 2.8). At the end of a session, the experimenter lets each subject roll a 20-sided die to determine the which round she gets paid for. The exchange rate between ED and Canadian dollars is $0.35 CAD per ED in COOP and $0.25 CAD per ED in COMP. These rates are chosen so that the ex ante expected payoff of the subjects are approximately $23 CAD for 60 to 75 minutes in the lab.

### 2.4 Results

21 sessions of the experiment with 136 subjects were conducted between June and October 2016 in SFU’s CRABE lab. 6 sessions (with 40 subjects) were in COOP, and 15 sessions (with 96 subjects) were in COMP. The number of observations in the COMP treatment is higher because decisions in this treatment exhibit higher variation. Within the COOP treatment, 20 subjects started with three messages in the public list and zero in the private list, while the other 20 subjects started with three messages in the private list and none in the public list.
Within the COMP treatment, 46 subjects started with no message in the public list while the other 50 subjects started with no message in the private list.

The average quiz scores were 8.40 in COOP and 8.23 in COMP, indicating that subjects had sufficient understanding of the games and the UI. Furthermore, the initial choices of 91% of the subjects in COOP were sent using a public message, and 76% of the subjects in COMP initially sent a private message. This result suggests that the subjects could by and large appreciate the strategic incentives underlying the use of the two types of messages.

In this experiment, I am interested in how pairs of message lists change over the course of 20 rounds, as they get modified and used by different pairs of subjects. Figure 2.9 shows the average number of words added, either by learning \( |E_{ij}| \) or by creating \( |N_{ij}| \), to each vocabulary over time, as well as the net linguistic differentiation \( |E_{ij}| - |N_{ij}| \).

In terms of the dynamics of learning, very similar trends are observed in both initial conditions of the COOP treatment. The number of words learned settles on a steady path near zero after the fifth round, despite the initial configurations of the message lists. However, there is a sustained, positive level of learning throughout a session in the COMP treatment. These observations are consistent with CM’s theory. Subjects in the COOP treatment only need one word for each action in the common list to be able to successfully coordinate on a payoff dominant equilibrium. Hence, after this condition is met, learning (as well as creating) should stop. On the other hand, in the COMP treatment, there is a constant incentive to stymie the possibility of the opponent hiding her intended action. Hence, subjects should always learn from their counterparts.

In COOP, there is no incentive for word creation. Indeed, this intuition is borne out by the data: despite some initial adjustments (possibly due to subjects discovering the incentive
Figure 2.9: Evolution of Message Lists over Time
structure of the game), the number of words created are by and large not statistically different from zero. In COMP, on the contrary, there is always a motive to create new words, since it affords a subject the ability to remain secretive in her choice of actions. As expected, we see that word creation in COMP is significantly greater than zero, and that the effort to create new words is sustained throughout a session.

From the bottom two panels of Figure 2.9, it is clear that the net linguistic differentiation in the COMP treatment is significantly higher than in the COOP treatment. To confirm, I perform a Mann-Whitney rank-sum test. In the experiment, an independent observation consists of the choices made by a pair of subjects over 20 rounds. Therefore, I use the average net linguistic differentiation (across pairs and rounds) to calculate the test statistic. The result is a rejection, with high confidence, of the hypothesis that net linguistic differentiations in the two treatments are the same ($z = -35.082$, with an associated $p$-value being virtually zero).\(^{12}\)

In addition, I perform regression analyses, the results of which are reported in Table 2.1. The dependent variable in the four regressions is net linguistic differentiation ($|E_{ij}^t| - |N_j^t|$). The first two columns contain estimates from simple OLS regressions. The third column contains estimates from a multi-level mixed effects linear regression, in which unobserved heterogeneities within each session and within each pair are controlled for. The coefficients in the fourth column are obtained from estimating a random-effects model that takes

![Table 2.1: Regression Analysis](image)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMP</td>
<td>$-0.837^{***}$</td>
<td>$-0.823^{***}$</td>
<td>$-0.822^{***}$</td>
<td>$-0.823^{***}$</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
<tr>
<td>Pairs/Session</td>
<td>$-0.024$</td>
<td>$-0.024$</td>
<td>$-0.028$</td>
<td>$-0.024$</td>
</tr>
<tr>
<td></td>
<td>(0.627)</td>
<td>(0.619)</td>
<td>(0.531)</td>
<td>(0.614)</td>
</tr>
<tr>
<td>Same $L_i^0$</td>
<td>0.130</td>
<td>0.138</td>
<td>0.130</td>
<td>0.130</td>
</tr>
<tr>
<td></td>
<td>(0.242)</td>
<td>(0.284)</td>
<td>(0.284)</td>
<td>(0.228)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.224</td>
<td>0.161</td>
<td>0.183</td>
<td>0.161</td>
</tr>
<tr>
<td></td>
<td>(0.521)</td>
<td>(0.604)</td>
<td>(0.581)</td>
<td>(0.598)</td>
</tr>
<tr>
<td>Observations</td>
<td>2720</td>
<td>2720</td>
<td>2720</td>
<td>2720</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.098</td>
<td>0.101</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dependent variable: $|E_{ij}^t| - |N_j^t|$

$p$-values in parentheses, standard error clustered at session level

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

\(^{12}\)The computation is performed in STATA using the `ranksum` command. The conclusion is robust to the choice of the unit of independent observations.
into account subject-level heterogeneity. Standard errors are clustered at the session level whenever applicable.\footnote{STATA commands used are \texttt{reg} for OLS, \texttt{xtmixed} for multi-level mixed effects, and \texttt{xtreg} for random-effects.}

Three independent variables are of particular interest. “COMP” is a dummy variable that equals 1 if and only if a subject is in the COMP treatment. The coefficient on this variable captures the effect of being in the COMP treatment on linguistic differentiation, holding other right-hand side variables constant. “Pair/Session” captures the number of pairs of subjects in a particular session. This is equal to the number of subjects in a session divided by 2. “Same $L_i^0$” is a dummy variable that equals 1 if and only if the subjects start out with the same initial dialects.

We see that the results are similar across the different specifications: The only significant determinant of net linguistic differentiation is whether subjects are in the COMP treatment. In other words, compared to subjects in the COOP treatment, those in COMP exhibit a stronger tendency to create more messages. More specifically, in light of the observation that vocabularies in the COOP treatment exhibit virtually no divergence, we can interpret the coefficient on COMP to mean that the average pair of vocabularies in the COMP treatment diverges at the rate of about 0.8 word per round.

2.5 Related Literature

Crawford (1998) surveyed experimental works on cheap talk games done in the late 1980s and the early 1990s. Of relevance to the present project is a paper by Blume et al. (2001), who observed in cheap talk games with partial common interest that the \textit{a priori} meanings of messages would get “recoded” during the course of the experiment.\footnote{A working paper version of the work by Blume et al. (2001) was cited in Crawford (1998), hence the time discrepancy.} More recently, Blume and Ortmann (2007) studied the effect of pre-play communication on coordination games with multiple Pareto-ranked equilibria. Messages with \textit{a priori} meanings were used in the communication. The authors found that pre-play communication had a marked effect on helping players coordinate on the Pareto dominant equilibrium. This finding is consistent with what occurs in my experiment.

2.6 Conclusion

The main prediction in CM’s paper is that languages become more numerous in environments where conflict is present. This assertion, albeit striking, would be hollow without a proper empirical validation. In light of the difficulty of carrying out a conventional econometric exercise, which would require panel data on linguistic diversity, this paper
resorts to the experimental approach. The lab environment affords the researcher a tremendous amount of control. For example, I am able to translate CM’s model directly into the lab setting and to randomize the assignment of subjects into different treatment groups. These measures help produce a cleanly estimated effect of strategic incentives on linguistic diversity.

The experimental evidence lends strong support to CM’s theory. In particular, I observe that, regardless of whether they start out to be the same, pairs of dialects in the competitive setting diverge in a statistically significant fashion, at a rate of 0.8 word per period. In contrast, dialects in the cooperative environment show little signs of divergence: they remain virtually unchanged if they are initially the same, or they converge eventually if they are initially distinct.

The experimental approach may prove useful for examining other aspects of linguistic diversity. For instance, agents in CM’s framework are assumed to be myopic for the sake of tractability. In experiments, such an assumption can be relaxed, thus introducing true dynamics into the evolution of linguistic diversity. Such an avenue of research is surely worth pursuing.
Chapter 3

The Evolutionary Foundation of the Preference for Surprise

3.1 Introduction

People exhibit preference for surprises, namely, the discovery of something unexpected. Spoiler alerts in movie reviews indicate that many people prefer some degree of uncertainty about the plots of the movies they are going to see. A tennis match with two comparably skilled players is usually more exciting than a match in which one player is almost sure to win (Ely et al., 2015). Close political elections garner more public attention and voter turnout than one-sided races (Geys, 2006). Humor is pleasant not least because surprise is a necessary ingredient (Hurley et al., 2011). These examples suggest that surprise, and the anticipation thereof, may be an important motivating factor for a wide range of seemingly unrelated activities. Admittedly, many other considerations—fascinating cinematography, dazzling athleticism, civic duty, and witty wordplay—also influence people’s behavior in the above cases. But it seems clear and plausible that people’s desire for surprise plays a role, too.¹

A key indication of the preference for surprise is people’s willingness to pay for it. Between 1995 and 2017, movies of the thriller/suspense genre have the fifth largest market share (8.37%) in the U.S. Many movies in the other top ranking genres, adventure (22.67%), comedy (21.05%), drama (16.43%), and horror (4.60%) also feature surprise as an important element.² Besides movies, people also spend considerable time and money on other entertainment activities such as sports events and casinos, whose consumption values are generated through information revelation. As Aguiar et al. (2013) observe, U.S. adults spend roughly 20% of their waking hours on non-income-generating entertainment activities.

¹My model also permits consideration of distaste of surprises. However, due to the messiness of algebra, I have not been able to derive an interpretable condition under which the disutility for surprise is evolutionarily optimal. I briefly discuss this aspect of the model in the end of Section 3.4.4.
In the aforementioned examples and numerous others, people care not just about the information *per se* but about the way information is revealed. Moreover, the information people demand is often non-instrumental: it is not obtained for the purpose of informing their decisions. For these reasons, this paper focuses on the preference for informational surprises. As Ely et al. (2015) point out, “the key feature of surprise is the ex post experience of a change in beliefs.” Hence to formalize the notion of surprise, I adopt the definition proposed by Itti and Baldi (2009), which measures surprise as the difference between a Bayesian individual’s prior and posterior beliefs after some event. Specifically, the surprise elicited by event $x \in X$, given a prior belief $f(\mu)$ about some state $\mu \in M$ and the Bayesian posterior $g(\mu | x)$, is given by

$$
\psi(g || f) = \int_{\mu \in M} \ln \left( \frac{g(\mu | x)}{f(\mu)} \right) g(\mu | x) d\mu.
$$

This is the Kullback-Leibler divergence of the prior from the posterior. This measure of surprise has a number of desirable properties. First, $\psi(g || f)$ is always non-negative, and equals zero if and only if $f = g$ almost everywhere. Thus $\psi$ quantifies the level of surprise an individual experiences due to an event $x$, be it the ending of a story, the result of a match, or the outcome of an election. Second, the Bayesian setting captures the subjective aspect of the experiences of surprises: individuals with different prior beliefs about the state may experience varying levels of surprises after the same event. Third, $\psi$ lends itself to a nice informatic interpretation: it represents the amount of additional information, in bits, contained in the posterior thanks to the observation of $x$. Therefore, the preference for surprise, thus measured, is a preference over the stochastic path of one’s beliefs.

Closely related to this formulation of the preference for surprise is the notion of curiosity. Loewenstein (1994) defines curiosity as the “information gap [between] what one knows and what one wants to know.” He further argues that a “curious individual is motivated to obtain the missing information”. Thus, from a behavioral perspective, the preference for surprise and curiosity both induce the same behavior of information collection. It is worth noting that Loewenstein (1994) also uses relative entropy to measure curiosity. Therefore, the preference for surprise can be easily reinterpreted as a preference for (non-instrumental) information resulting from curiosity.

This paper proposes an evolutionary foundation of such a preference. I argue that the taste for surprise is evolutionarily optimal in an environment where i) signals about the state become more informative over time at a fast rate, but ii) individuals have a form a cognitive inertia, i.e. they are predisposed to believe that the informativeness of the signals remains unchanged. In other words, the evolutionary purpose of the preference for surprise is to motivate learning and experimentation in an environment where individuals consistently underestimate the benefits to information acquisition.
To fix ideas, consider individuals living in a hunter-gatherer society. Suppose these individuals are originally of different “surprise types”, i.e. they exhibit varying degrees of preference for surprise. The surprise types are determined exogenously: think of this as the random mutation that occurs during the process of evolution to the biological mechanism governing a person’s emotional reactions to unanticipated events.\(^3\) Let \(\sigma\) denote an individual’s surprise type; that is, \(\sigma\) measures the intensity of one’s taste for surprise. Then \(\sigma = 0\) would suggest indifference to surprises; a large positive value would indicate a strong preference for surprises; and a negative \(\sigma\) signifies aversion to surprises. If we assume that surprise types are inherited by offspring, then the most fitting surprise type in a given environment, i.e. the \(\sigma\) that induces the maximum fitness in every period, will eventually dominate the population. Put figuratively, it is as if the process of evolution is maximizing individual fitness by selecting the right surprise type.

In my model, an individual makes two decisions every period. First, he decides how much effort to spend on collecting information about some unknown state. For instance, the state could be an unknown disease he has caught, and the information collected would be pieces of advice from his fellow hunter-gatherers. The accuracy of the advice reflects the level of general medical knowledge in the community, and this level is increasing over time. Second, based on the information gathered, the individual chooses an action that bears reproductive consequences, e.g. taking a course of treatment to cure the disease. His fitness, measured as quality-adjusted offspring, will be determined jointly by the state and the actions he chooses. In light of my definition of surprise, different values of \(\sigma\) imply different decisions; in particular, an individual who likes surprises (\(\sigma > 0\)) will tend to gather more information than an individual who is indifferent to surprises (\(\sigma = 0\)). Since we assume that individuals are prone to underestimate the informativeness of the advice they get, the ones with \(\sigma = 0\) will in general gather a suboptimal amount information. As a result, the surprise-neutral types will be disfavored by evolution.

Crucial to this conclusion is the assumption of the individual’s inclination to underestimate the usefulness of information. If we think of the advice the individual gathers as a technology of discovering the state, and that this technology improves over time, then the assumption on the individual’s predisposition would be the same as requiring that the individual adopt new technologies with a time lag, a well-known empirical regularity (see, for example, Rosenberg, 1976; Mansfield, 1968). Studies in psychology and management science have also identified persistence in beliefs—the so-called cognitive inertia—as a cause of the failure to adequately adjust to the changes in the environment (Ross and Anderson, 1982; Hodgkinson, 1997; Tripsas and Gavetti, 2000).

\(^3\)One such candidate is the neurotransmitter dopamine, which is hypothesized to encode “reward prediction error”, that is, the degree to which a reward is surprising. See Caplin and Dean (2008) and Caplin et al. (2010) for details.
The rest of the paper is organized as follows. Section 3.2 contains a brief review of related previous studies. Section 3.3 sets up the model. In Section 3.4, I present main results of the paper. As a benchmark, I first consider in Section 3.4.1 the behavior of an individual who is indifferent to surprises. Then in Section 3.4.2 I derive the first best behavior without imposing the assumption of cognitive inertia. Section 3.4.3 considers the utility maximization problem faced by an individual with utility for surprise. Lastly, in Section 3.4.4 I derive conditions under which the utility for surprise is evolutionarily optimal. Section 3.5 concludes.

3.2 Related Literature

3.2.1 Evolution as Fitness Maximization: The Principal-Agent Metaphor

As alluded to earlier, the process of natural selection can be viewed, via a metaphorical lens, as evolution solving a fitness maximization problem. Robson (2001a,b) advocates considering the problem within a principal-agent framework, where the principal is the blind force of evolution and the agent is the individual. While both the principal and the agent ultimately seek to maximize the latter’s fitness, they do so in different manners. The agent has a utility function, and is assumed to be a utility maximizer. The agent’s utility is in general positively correlated with his fitness, so he maximizes fitness indirectly. A key friction in this type of problem is that the agent does not have access to the same information that the principal does, and the principal cannot directly communicate her information to the agent. The information available to the principal is accumulated from the long history of evolution, and it is biologically infeasible to encode such a vast amount of data in its entirety into the human brain. Hence, in situations where the agent lacks pertinent information (but the principal does), his decisions would lead to suboptimal fitness. The principal has the ability to guide the agent’s decisions by designing his utility function, taking into account the agent’s utility maximizing behavior and his lack of information. The principal-agent approach has been employed in previous work to provide evolutionary justifications of economic behaviors. For instance, Rayo and Becker (2007) use this approach to explain why the experience of happiness, or hedonic utility, tends to be influenced by factors such as past successes, peer comparisons, prior expectations, etc. They argue that evolution favors a utility function that measures happiness in relative terms. Rayo and Robson (2014) demonstrate that preference over outcomes other than offspring could serve an evolutionary purpose when a two-sided information asymmetry exists between Nature and the individual. The authors theorize that the preference for intermediate outcomes evolves as a result of the individual’s ignorance of their significance in determining fitness.

4Throughout the paper, I also use the female pronoun, “she”, to refer to the principal, as it is customary in English to refer to nature as Mother Nature, and the male pronoun “he” to refer to the agent.
3.2.2 Preference for Surprise

This paper is partly inspired by Ely et al. (2015), who rightly argue that “the key feature of surprise is the ex post experience of a change in beliefs.” The authors posit a utility function for surprise (and suspense), and study optimal information disclosure policies under such a preference. Their baseline model features a finite state space, and the utility for surprise is measured by the Euclidean distance between the prior and posterior. In this paper, I provide an evolutionary explanation for why a preference for surprise has evolved. Moreover, the state space in my model is a continuum, and I measure surprise differently (using the Kullback-Leibler divergence).

On the empirical front, the neurotransmitter dopamine has been proposed as the biological manifestation of the preference for surprise. A study by Mirenowicz and Schultz (1994) on monkeys reveals that the dopaminergic response to a juice “reward” depends largely on whether the reward is anticipated. In the experiment, the researchers measure the monkeys’ dopaminergic responses as they learn to associate a tone with the receipt of a fruit juice moments after hearing the tone. It is found that, before acquiring the association between the tone and the juice, a monkey’s dopamine neurons fire in response to the juice but not the tone; however, after the association is learned, the dopamine neurons react to the tone but not the juice. A similar pattern is observed in human subjects in fMRI studies (Montague and Berns, 2002). These findings lead neuroscientists to believe that dopamine encodes the reward prediction error, that is, the extent to which a reward is surprising, as opposed to the reward itself. Caplin and Dean (2008) provide an axiomatic foundation for a dopaminergic release function consistent with the reward prediction error hypothesis. The main axioms of their paper are tested by Caplin et al. (2010), and the data show varying degrees of support.

3.3 Model

In this paper, I adopt the language of the principal-agent metaphor in developing a formal framework to analyze the evolution of the preference for surprise. The readers should bear in mind that the adoption of the metaphorical speech is merely for the ease of exposition, and that I do not attribute consciousness to Nature or evolution.

Environment Let \( \{ \tilde{\mu}_t : t = 1, 2, \ldots \} \) be a collection of real-valued random variables defined on the same probability space \( (\Omega, \mathcal{F}, P) \), where \( \Omega \) is an underlying sample space with \( \sigma \)-field \( \mathcal{F} \) and a probability measure \( P \). Thus \( \{ \tilde{\mu}_t \} \) is a stochastic process with state space \( \mathbb{R} \) and parameter set \( \{1, 2, \ldots\} \). I refer to \( \mu_t \), the realization of \( \tilde{\mu}_t \), as the state in period \( t \). For each \( t \geq 1 \), assume that \( \tilde{\mu}_t \) is normally distributed with mean \( \mu_{t-1} \) and variance \( s_t \). In the interest of simplicity, I assume that \( s_t = s_{t+1} = s \), i.e. the variance of the state
distribution is constant over time. Further assume that $\mathbb{E}(\tilde{\mu}_{t+1} | \mu_1, \ldots, \mu_t) = \mu_t$ for all $t \geq 1$. Thus the stochastic process $\{\tilde{\mu}_t\}$ is a martingale.

In each period, individuals can observe signals about the state in that period. For a particular realized state $\mu_t$, a signal is given by $x_t = \mu_t + \epsilon_t$, where $\epsilon_t$ is the realization of a normally distributed random variable $\tilde{\epsilon}_t$ with mean zero and time dependent variance $\nu_t$.\(^5\) I interpret $\epsilon_t$ as the error contained in the signal. Thus the signals are unbiased about $\mu_t$ and $x_t \sim \mathcal{N}(\mu_t, \nu_t)$. The time dependent variances form a sequence of real numbers $\{\nu_t : t = 1, 2, \ldots\}$, which I assume can be described by a parameterized function.

**Assumption 3.1.** For all $t \geq 1, \nu_t > \nu_{t+1}$.

If we interpret $\nu_t$ as the technology of discovering the state at time $t$, with smaller $\nu_t$ representing more advanced technology, then Assumption 3.1 says that this technology gets better and better every period. In reference to the medical example in the introduction, Assumption 3.1 says that the general medical knowledge of the society is improving over time. Further, if such an improvement of knowledge (or technology in general) is governed by some variant of Moore’s law, then we can write $\nu_{t+1} = \alpha \nu_t$, where $\alpha \in (0, 1)$ is the rate of technological progress.\(^6\)

**Agent** The agent and his offspring are modeled in an overlapping generation fashion. Each generation $t \in \mathbb{N}$ is populated by a continuum of agents who live for two periods, youth and adulthood. In each period $t$, the youth of generation $t+1$ and the adults of generation $t$ coexist. Thus, I use the notation $t$ to index both a time period and the generation whose adults live in that period.

An adult agent (of generation $t$) in period $t$ decides sequentially on two things. First, he chooses how much information, $n_t \in \mathbb{R}_+$, to gather. Loosely speaking, $n_t$ is the number of signals sampled from the signal distribution $\mathcal{N}(\mu_t, \nu_t)$. While sample sizes generally take integer values, here I treat $n_t$ as a continuous quantity so that I could use differentiation to find the optimal $n_t$. Since in my model all distributions are Gaussian, assuming a continuous $n_t$ raises no technical issues. To interpret a non-integer $n_t$, we can think of it as representing the effort put into collecting one signal from a distribution whose variance decreases continuously in $n_t$; and at integer values of $n_t$, the variance of this signal is equal to the sample variance of $n_t$ independent observations from $\mathcal{N}(\mu_t, \nu_t)$. For this reason, I keep the interpretation of $n_t$ as a generalized sample size for the rest of this paper. Based on the information collected, the agent then takes an action $a_t \in \mathbb{R}$. The fitness of an adult

\(^5\)If multiple signals are sampled, I use $x_{t,j}$ (or simply $x_j$, when the time subscript is suppressed) to indicate the $j^{th}$ signal sampled in period $t$.

\(^6\)Moore’s law (Moore, 1965) is the observation that the number of transistors in an integrated circuit doubles approximately every two years. Since the number of transistors on a chip is closely related to the overall processing power for computers, Moore’s law can be understood as a description of the rate of technological change.
agent in a given period $t$ is determined jointly by the sample size $n_t$, the action $a_t$, and the state $\mu_t$ via a biological production function

$$\phi(n_t, a_t; \mu_t) = Q(\mu_t) - (a_t - \mu_t)^2 - n_t c.$$  \hspace{1cm} (3.1)

An agent’s fitness represented in (3.1) is measured by the number of quality-adjusted offspring, where $Q(\mu_t) > 0$ for all $\mu_t$ is the maximally attainable number of quality-adjusted offspring determined by a particular state $\mu_t$; and $c$ is the marginal cost of information acquisition. The term $Q(\mu_t)$ is introduced so that we do not have to deal with the interpretation of negative fitness in most cases (assume that the expectation of $\mu_t$ is far enough away from zero). As I show in Section C.1, $Q(\mu_t)$ does not affect the agent’s optimal decisions.

A young agent (of generation $t + 1$) living in period $t$ has costless access to the signals observed by everyone in the elder generation in that period. Specifically, if $x_i^t = (x_{i,1}^t, \ldots, x_{i,n_i^t}^t)$ denotes the set of $n_i^t$ signals observed by an adult agent $i$ in period $t$, then a young agent in that period observes the set $\{x_i^t\}_i$ for every adult agent $i$ living in that same period. Since there is a continuum of generation-$t$ adults, the young agent can infer the state $\mu_t$ and the variance of the signal distribution $\nu_t$ from the observations, due to the law of large numbers. As the young agent enters a new period as an adult, his prior about the state and the signals in $t + 1$ will depend on the inferred parameters, $\mu_t$ and $\nu_t$, as well as $s$, which is assumed to be known by all agents.\footnote{Assume for simplicity that all young agents survive into adulthood. The assumption that young agents are able to make the inference of the parameters while the adults are not is made in part for mathematical convenience. Nevertheless, the assumption can also be justified by considering the urgency of the adults’ decisions. In particular, suppose the adults need to choose an $a_t$ by a certain deadline and do not have sufficient time to pool all the information collected. On the other hand, young agents do not face such an urgency. For one thing, they live longer than the adults; for another, the only thing young agents do is to observe. Therefore, it is reasonable to assume that young agents are better able to infer the relevant parameters than their older counterparts.}

To economize on notation, I henceforth suppress all time subscripts, using $\mu$ in place of $\mu_t$ and $\mu'$ in place of $\mu_{t-1}$ for the realized states in the respective periods. Moreover, I use $\tilde{\mu}$ to denote the unrealized state in period $t$. The same notational convention applies to variables other than $\mu$.

The agents of all generations know that $\mu$ follows a martingale process. However, they do not know $\alpha$, the rate of technological advancement. The prior of a time-$t$ adult agent about the current state is modeled as a normal distribution $\mathcal{N}(\mu', s)$ with mean $\mu'$ and variance $s$. Additionally, I make an important assumption about the agent’s belief.

**Assumption 3.2.** A time-$t$ adult agent believes that the signal distribution at time $t$ has variance $\nu'$.

Thus, the adult agent believes that the signals observed in the current period are as informative as the ones in his parent’s time. In other words, the adult agent underestimates the informativeness of the signals. There are several justifications.
First, we can think of this assumption as a result of anchoring, a widely observed behavioral bias. In short, anchoring occurs when people rely disproportionately on pieces of information that arrive first (an implicit “reference point”) during their belief updating process (Tversky and Kahneman, 1974). Since the adult agent’s belief about the precision of the signals is formed primarily by observing the signal realizations in the previous generation, Assumption 3.2 would be consistent with an extreme form of anchoring whereby a time-\(t\) adult agent relies exclusively on data from \(t-1\) to estimate the signal precision.

Alternatively, we can think of Assumption 3.2 as describing a time lag in the full utilization of new technology. When a piece of new technology becomes available, there is usually a significant delay before it is fully utilized. For example, the link between smoking and ill-health was discovered 54 years before cigarette advertisements were banned in the U.K. (Hanney et al., 2015). Screening technology for abdominal aortic aneurysm was discovered 26 years before it became publicly available (Hanney et al., 2015). It took 39 years for land-line telephones to penetrate 40% of the U.S. market; even mobile phones, allegedly the fastest spreading technology in human history, took 10 years to reach 40% of U.S. households (McGrath, 2013). These are merely adoption rates. Technology can be adopted but not fully utilized. Since \(\nu'\) and \(\nu\) can be construed as measuring the technology of discovering the state in \(t-1\) and \(t\), respectively, the agent’s predisposition to use \(\nu'\) in period \(t\) represents his under-utilization of the technology at \(t\).

A third justification for Assumption 3.2 is through the agent’s relative informational disadvantage. An adult agent only has access to data from one period ago, but the principal has access to all data since the beginning of time. Hence if \(\nu = a\nu'\), or if the sequence of \(\nu\) is generated by some other simple parametric function, the principal can estimate the parameter \(a\), and thus predict \(\nu\) with much higher accuracy than the agent could. Assumption 3.2 captures this comparative disadvantage of the agent by not letting him use the correct variance of the signal distribution.

Lastly, this assumption is roughly consistent with the notions of cognitive inertia in management science (Hodgkinson, 1997; Tripsas and Gavetti, 2000) and belief perseverance in psychology (Ross and Anderson, 1982), both of which refer to the phenomenon that beliefs tend to endure once they are formed.

The information collection is modeled as follows. The adult agent can choose to gather \(n\) independent signals about \(\mu\) for a total cost of \(nc\). The cost is measured in terms of the loss of potential offspring, as time and energy spent on gathering information means those resources cannot be used on reproductive and child-rearing activities. Let \(x_1, \ldots, x_n\) be the outcomes from sampling, and let \(\bar{x} = \frac{1}{n} \sum x_j\) be the sample average. Thanks to the conjugacy of the normal distribution, the posterior about the current state after sampling is also a normal distribution \(\mathcal{N}(\beta_m, \beta_s)\) with mean

\[
\beta_m = \frac{\mu'\eta' + n\bar{x}}{\eta' + n}
\] (3.2)
and variance
\[ \beta_s = \frac{s \eta'}{\eta' + n}, \]
where \( \eta' = \nu'/s. \)

Observe in (3.2) that the posterior mean can be written as a convex combination between the prior mean and the sample average, with weight \( \eta' / (\eta' + n) \). Hence \( \eta' \) can be interpreted as the equivalent number of sample observations that describes the amount of information implicit in the prior distribution. A large \( \eta' \) implies less precision of the signal distribution relative to the prior, and so the prior mean weighs more heavily in \( \beta_m \).

**Principal** The principal knows that the state follows a martingale process \( \{\tilde{\mu}_t\} \). She also knows the function that generates \( \nu \) (i.e. the parameter \( \alpha \)), but she cannot communicate this to the agent. Given this knowledge, her objective is to maximize the expected fitness of the agent, taking into account how the agent forms his prior and posterior, and the cost of acquiring information. The principal does so by designing a utility function for the agent, which the agent then maximizes in his adulthood.

Hence, at \( t = 0 \), the principal chooses a utility function for the agent, anticipating that the agent would maximize this function in his adulthood. Then within any \( t \geq 1 \), and where \( f(\cdot | y, z) \) denotes the density of a normal distribution with mean \( y \) and variance \( z \), the following events occur:

i) adult agent forms prior \( f(\mu | \mu', s) \) about the state;
ii) \( \mu, \nu \) determined but not yet revealed to the agent;
iii) adult agent chooses the sample size \( n \);
iv) adult agent observes signals \( x_1, \ldots, x_n \);
v) adult agent forms posterior \( f(\mu | \beta_m, \beta_s) \) about the state;
vi) adult agent chooses reproductive decision \( a \);
vii) fitness of adult agent, \( \phi = \phi(n, a; \mu) \), realizes;
viii) young agent observes the signals \( \{x'_i\}_1 \) collected by every adult agent \( i \) in \( t \).

### 3.4 Optimal Utility Function for the Agent

#### 3.4.1 Utility as Fitness: Agent’s Problem

If the agent’s utility function is simply his fitness, then the agent would solve the following problem:

\[
\max_{n \in \mathbb{R}_+} \left[ \mathbb{E}_X \left( \max_{a \in \mathbb{R}} \int_{\mathbb{R}} (Q(\mu) - (a - \mu)^2) f(\mu | \beta_m, \beta_s) d\mu \right) - nc \right].
\]
Given that the prior and the signal distributions are both normal, the marginal distribution of \( \bar{x} \), the sample average of \( n \) observations, is also normally distributed.\(^{10}\)

**Lemma 3.1.** There exists a solution to (3.4), given by

\[
a^A(n^A) = \beta_m = \frac{\mu'\eta' + n^A\bar{x}}{\eta' + n^A},
\]

which is the expected value of \( \mu \) under the posterior, and

\[
n^A = \left[ \frac{\nu'}{c} \right]^{\frac{1}{2}} - \frac{\nu'}{s},
\]

provided that \( n^A \) is non-negative; otherwise \( n^A = 0 \).

The proof is in Section C.1.

### 3.4.2 First Best Solutions

Suppose the principal were to decide for the agent in each period. She would utilize her knowledge about the parameter \( a \) in her decision making. Specifically, the principal’s prior about the current state would be a normal distribution with mean \( \mu' \) and variance \( s \), and she knows that the signal distribution has variance \( \nu \), instead of \( \nu' \) which is believed by the agent. Consequently, after obtaining a sample of size \( n \) with sample average \( \bar{x} \), the principal’s posterior is \( \mathcal{N}(\xi_m, \xi_s) \) with mean

\[
\xi_m = \frac{\mu'\eta + n\bar{x}}{\eta + n},
\]

and variance

\[
\xi_s = \frac{s\eta}{\eta + n},
\]

where \( \eta = \nu/s \). The principal then faces the same problem in (3.4), except with \( \beta_m, \beta_s \) replaced by \( \xi_m, \xi_s \), or essentially, with \( \nu' \) replaced by \( \nu \).

**Lemma 3.2.** To maximize fitness, the principal chooses

\[
a^*(n^*) = \xi_m = \frac{\mu'\eta + n^*\bar{x}}{\eta + n^*},
\]

and

\[
n^* = \left[ \frac{\nu}{c} \right]^{\frac{1}{2}} - \frac{\nu}{s},
\]

provided that \( n^* \) is non-negative; otherwise \( n^* = 0 \).

\(^{10}\)Pratt et al. (1995) show that \( \bar{x} \sim \mathcal{N}(\mu', \frac{s+V}{n}) \), where \( V \) is the perceived variance of the signal distribution. Therefore, \( V = \nu' \) for the agent, and \( V = \nu \) for the principal.
The proof is similar to that of Lemma 3.1. From (3.6) and (3.10) we see that the optimal number of signals depends crucially on the variance of the signal distribution (see Figure 3.1 for an illustration). In particular, if the variance of the signal distribution is too large relative to the variance of the prior, then there is no point sampling information from it because the signals contain too much noise. Since the agent is predisposed to believe, falsely, that the signal distribution has variance $\nu'$, his decisions would be suboptimal in general.

3.4.3 Agent’s Problem with Utility for Surprise

I measure the agent’s utility for surprise by the level of surprise he experiences.\(^\text{11}\) Thus, the utility for surprise is given by

$$U_{\text{surp}}(n; \bar{x}) = \int_{\mathbb{R}} \ln \left( \frac{f(\mu | \beta_m, \beta_s)}{f(\mu | \mu', s)} \right) f(\mu | \beta_m, \beta_s) d\mu.$$  

This utility represents the amount of additional information (relative to the prior) contained in the posterior due to the observations of signals $x_1, \ldots, x_n$. It is always non-negative, and equals zero if and only if the prior and posterior are the same. I assume that $U_{\text{surp}}$ is linearly separable in the agent’s utility function. Then the problem faced by the agent in a typical

\(^{11}\)I could, as is in Ely et al. (2015), assume that utility for surprise is an increasing and concave function of the level of surprise. Doing so does not alter the results qualitatively. Hence, in the interest of simplicity, I maintain the current functional form of the utility for surprise.
\[
\max_{n \in \mathbb{R}} \left\{ \sigma \mathbb{E}_x \left[ \int_\mathbb{R} \ln \left( \frac{f(\mu | \beta_m, \beta_s)}{f(\mu | \mu', s)} \right) f(\mu | \beta_m, \beta_s) d\mu \right] + \mathbb{E}_x \left[ \max_{a \in \mathbb{R}} \int_\mathbb{R} (Q(\mu) - (a - \mu)^2) f(\mu | \beta_m, \beta_s) d\mu \right] - nc \right\}
\]

(3.11)

where \(\sigma\) is selected by the principal at \(t = 0\).

**Lemma 3.3.** Let \(a^{A^*}\) and \(n^{A^*}\) be the solution to (3.11). Then,

\[
a^{A^*}(n^{A^*}) = \frac{\mu' \eta' + n^{A^*} \bar{x}}{\eta' + n^{A^*}},
\]

and

\[
n^{A^*}(\sigma) = \left[ \frac{\sigma^2}{4} + 4sc\eta' \right]^{\frac{1}{2}} + \frac{\sigma}{4c} - \eta',
\]

(3.12)

provided that \(n^{A^*}\) is non-negative; otherwise \(n^{A^*} = 0\).

The proof is in **Section C.2**. Observe that, at \(\sigma = 0\), \(n^{A^*}(0) = n^A\). Therefore, \(\sigma\) captures how the utility from surprise affects the agent’s choice of \(n\), which indirectly influences the decision \(a\). Moreover, note that the function \(n^{A^*}(\sigma)\) is onto; that is, for every \(\hat{n} \in \mathbb{R}_+\), there exists a unique \(\hat{\sigma} \in \mathbb{R}\) such that \(n^{A^*}(\hat{\sigma}) = \hat{n}\). This can be seen by solving (3.12) for \(\sigma\) as a function of \(n\):

\[
\sigma(n) = \frac{2c(\eta' + n)^2 - 2s\eta'}{\eta' + n}.
\]

Since \(n^{A^*}(\cdot)\) is also strictly increasing, as is evident from (3.12), I conclude that \(n^{A^*}\) is a bijection. Lastly, note that \(n^{A^*}(\sigma) \to \infty\) as \(\sigma \to \infty\) and vice versa.

### 3.4.4 Evolutionary Optimality of the Utility for Surprise

To maximize the agent’s actual expected fitness subject to the agent’s cognitive inertia and utility maximizing behavior, the principal solves the following problem:

\[
\max_{\sigma} \mathbb{E}_x \left( \int_\mathbb{R} \left( (Q(\mu) - (a^{A^*}(\sigma) - \mu)^2 \right) f(\mu | \xi_m, \xi_s) d\mu \right) - n^{A^*}(\sigma)c,
\]

(3.13)

where \(\sigma\) influences \(n^{A^*}\) and indirectly \(a^{A^*}\), since the latter is a function of \(n^{A^*}\). Since \(n^{A^*}\) is a bijection, we can solve (3.13) by directly choosing \(n^{A^*}\). Similar to the derivation of (C.3) in **Section C.1**, we can rewrite (3.13) as

\[
\max_n -\mathbb{E}_x \left[ \left( \frac{\mu' \eta' + n \bar{x}}{\eta' + n} - \frac{\mu' \eta + n \bar{x}}{\eta + n} \right)^2 \right] - \frac{\nu}{\eta + n} - nc,
\]

59
or equivalently,

\[
\min_n \left( \frac{\eta' - \eta}{\eta' + n} - \frac{\eta}{\eta + n} \right)^2 \left( \frac{\nu'}{n} \right) + \frac{\nu}{\eta + n} + nc. \quad (3.14)
\]

Observe that without the terms in the square bracket, the solution to (3.14) would be the same as the one in (3.10). Let \( \hat{n} \) denote the solution to (3.14). Then the Kuhn-Tucker condition characterizing \( \hat{n} \) is

\[
\frac{s(\eta' - \eta)^2(\eta\eta' - \eta\hat{n} - 2\hat{n}^2)}{(\eta' + \hat{n})^3(\eta + \hat{n})^2} - \frac{s\eta}{(\eta + \hat{n})^2} + c \geq 0, \quad (3.15)
\]

where a strict inequality implies \( \hat{n} = 0 \). Section C.3 provides a brief derivation and shows that (3.14) is strictly convex on \([0, \infty)\) and that the left-hand side of (3.15) is positive as \( \hat{n} \to \infty \). Hence, \( \hat{n} \) is the unique solution that solves (3.14). Observe that \( \hat{n} = n^* = n^A \) when \( \eta' = \eta \). This reinforces the intuition that if the agent does not suffer from cognitive inertia, then surprise need not play a role in the agent’s utility function.

**Theorem 3.1.** Let Assumptions 1 and 2 hold. Then \( \hat{n} > n^A \) if and only if the parameters \( s, c, \eta' \) and \( \eta \) are such that

\[
(\eta' - \eta)^2 \left[ 2\eta\eta' - \eta \left( \frac{S}{c} \eta' \right)^{1/2} - 2 \left( \left( \frac{S}{c} \eta' \right)^{1/2} - \eta' \right) \right] - \eta \left( \frac{S}{c} \eta' \right)^{3/2} + \left( \frac{S}{c} \right)^{1/2} \left( \eta' - \eta - \left( \frac{S}{c} \eta' \right)^{1/2} \right)^2 < 0. \quad (3.16)
\]

The proof is in Section C.4. Figure 3.2 visually compares the problems in (3.4) and (3.14), for a particular set of parameter values. In the figure, the dashed (blue) curve plots (3.14), the dash-dot (red) and the solid (black) curves are respectively the objective functions used to solve for \( n^A \) and \( n^* \). The minimum of these three curves correspond to the optimal \( n \)'s.

**Proposition 3.1.** The utility for surprise is evolutionarily optimal, i.e. \( \sigma^+ > 0 \), if

\[
\nu' - \nu \geq 3\sqrt{\frac{\nu'}{c}}.
\]

This follows from (3.16), together with the observation that \( n^{A^*}(0) = n^A \) and \( n^{A^*}(\sigma) \) is strictly increasing in \( \sigma \). See Section C.5 for details. Thus Proposition 3.1 says that if the variance of the signal distribution decreases by a sufficiently large amount across periods, then it is evolutionarily optimal to let the agent derive pleasure from surprises.

---

\[12\] The parameter values used to produce this graph are as follows: \( s = 1, \eta' = \nu' = 15, \eta = \nu = 1.5, c = 0.05. \]
Remark 3.1. The current framework also permits the analysis of unpleasant surprises, which is evolutionarily optimal when \( \hat{n} < n_A \). This occurs if the inequality in (3.16) is reversed. While I have not solved for a condition that is interpretable, my conjecture is that the disutility of surprise is evolutionarily optimal when \( \nu' \) is sufficiently smaller than \( \nu \).

### 3.5 Discussion

I have presented a theoretical argument for why people possess a taste for surprises. I show that the preference for surprise is evolutionarily optimal when an inertial force prevents individuals from collecting an optimal amount of useful information. The literature on the diffusion rate of technological innovation offers many reasons for why such an inertial force exists. While the utility for surprise is psychological and does not directly relate to fitness, such a utility enhances fitness in that it reduces the effective cost of learning, which leads to more informed decisions that bear consequences for fitness.
Bibliography


Appendix A

Supplementary Materials for Chapter 1

A.1 Justification for the Reduced Form Expected Payoffs

In this section we provide a justification for the two reduced form expected payoffs, (1.2) and (1.3), described in Section 1.3.3. In particular, we show how communication in games of cooperative and competitive natures can lead to the specific forms of these payoff functions.

Our argument is developed in a framework of two-player extensive form games, whose payoff structure is of either cooperative or competitive nature. The game begins with nature choosing a state and one randomly selected player to observe it. This player, referred to as the sender, then decides whether to transmit a message regarding the state. If a message is sent, then the other group, called the receiver, has an opportunity to choose an action based on the sender’s message.

Each group is endowed with a set of messages that refer to a subset of the states. Elements in the intersection of the two spaces are understood by both groups, while those unique to one group’s message space are understood only by that group. We derive a group’s expected payoffs in the subgame perfect equilibria, and show that they are related to the sizes of intersection and relative complement of the two message spaces in the way specified by (1.2) and (1.3).

A.1.1 General Environment

Suppose at the beginning of each period $t$, a two dimensional state $(\tau, a)$ is drawn to determine a game $\Gamma^t(\tau, a)$ played by groups $i$ and $j$. The parameter $\tau \in \{\text{coop, comp}\}$ determines the strategic nature of the interaction, and $a \in A = [0, 1]$ determines the optimal action(s) for each group. The two parameters are drawn independently of each other: $\tau = \text{coop}$ with probability $\frac{p}{p+q}$ and $\tau = \text{comp}$ with probability $\frac{q}{p+q}$; $a$ is drawn uniformly.

While we interpret a group’s message space to be its dialect, the framework proposed in this section does not integrate perfectly with the model presented in the paper. Specifically, in this section, the measure of the intersection of the message spaces is assumed to be smaller than the measure of the action space. For this to hold every period in the main sections of the paper, the action space would have to grow exogenously every period at a rate no smaller than the growth rate of the intersections.
from $A$. Let $\sigma_k$, $k = i, j$, denote group $k$’s action in $\Gamma^i(\tau, a)$. The payoffs in $\Gamma^i(\text{coop}, a)$ are

$$\pi_i(\sigma_i, \sigma_j) = \pi_j(\sigma_i, \sigma_j) = \begin{cases} 1 & \text{if } \sigma_i = \sigma_j = a \\ 0 & \text{otherwise,} \end{cases} \quad \forall a \in A; \quad (A.1)$$

and the payoffs in $\Gamma^i(\text{comp}, a)$ are

$$\pi_i(\sigma_i, \sigma_j) = \begin{cases} 2 & \text{if } \sigma_i = a \neq \sigma_j \\ -2 & \text{if } \sigma_i = \sigma_j = a \end{cases} \text{ and } \pi_j(\sigma_i, \sigma_j) = -\pi_i(\sigma_i, \sigma_j), \quad \forall a \in A. \quad (A.2)$$

Let $M_k = [x_k, y_k]$, where $0 \leq x_k < y_k \leq 1$ and $k = i, j$, be the set of messages that group $k$ can use to refer to actions in $A$. Note that $M_i \cap M_j$ need not be empty, and that $M_i$ and $M_j$ need not have the same size. Each message $m \in M_k$ refers to one and only one action $a \in A$. More precisely, let $\mu_k : M_k \to A$ be a measure-preserving injection that maps elements in $M_k$ to those in $A$. Thus, the meaning (or referent) of a message $m \in M_k$ is given by $\mu_k(m)$. Further, let $\mu_i, \mu_j$ satisfy the property that $\mu_i(m) = \mu_j(m)$ for all $m \in M_i \cap M_j$, so that messages in the intersection of the two message spaces refer to the same actions. Denote by $m_i^a$ a message sent by $k$ that refers to action $a$; thus $m_i^a = \mu_k^{-1}(a)$. We say that group $j$ understands a message $m_i$ sent by group $i$ if and only if $m_i \in M_i \cap M_j$.\textsuperscript{2}

Let $\tilde{\Gamma}^i$ be an extensive form game based on $\Gamma^i(\tau, a)$:

i) With equal probability, nature chooses either $i$ or $j$ to be the sender.

ii) The sender, say $i$, observes the states $(\tau, a)$.

- If the state $a$ is such that there exists no $m_i \in M_i$ with $\mu_i(m_i) = a$, the game ends and both $i$ and $j$ get zero payoff.
- If there exists $m_i \in M_i$ with $\mu_i(m_i) = a$, then $i$ decides whether to engage the receiver, $j$, in $\Gamma^i(\tau, a)$.
  - If $i$ chooses not to engage, then the game ends with both groups getting zero payoff.
  - If $i$ chooses to engage, then it sends $m_i^a$ and chooses $\sigma_i = a$ in $\Gamma^i(\tau, a)$. Both $\tau$ and $m_i^a$ are then observable to $j$.

iii) Based on $\tau$ and $m_i^a$, $j$ chooses an action $\sigma_j \in A$.

iv) The payoffs are determined by $\sigma_i$ and $\sigma_j$ according to (A.1) and (A.2).

The game $\tilde{\Gamma}^i$ can be interpreted as follows. The leader of group $i$ discovers an interaction opportunity characterized by $(\tau, a)$ that would involve another group $j$. Group $i$'s optimal payoff is a function of $a$. If group $i$ does not have the requisite language (i.e. $m_i^a$) to communicate $a$ among its members, then it cannot take advantage of this opportunity. Hence there is no change in either group’s payoff. Suppose $i$ has a message $m_i^a$ that refers to $a$. The the leader of $i$ can still decide whether to take advantage of the opportunity based on the anticipated response of $j$. If she decides not to, then there is no change in either group’s payoff. If she decides to take the opportunity, then she sends $m_i^a$, which serves as a coordination signal for members of group $i$. Group $j$ can also observes $m_i^a$ with some probability, and for simplicity, we assume this probability to be 1. Based on whether $j$ understands $m_i^a$ and the observed nature of the interaction, $j$ responds (optimally).

\textsuperscript{2}We assume that messages in $M_i \cap M_j$ is common knowledge.
In the next two subsections, we focus on the case where the sender has a message for the drawn state \( a \), and analyze a group’s expected payoff in a subgame perfect equilibrium (SPE).

### A.1.2 Cooperative Interaction

Suppose \( \tau = \text{coop} \), and \( i \) is the sender. The following is an SPE of \( \Gamma_t \):

- \( i \) always engages, sends \( m_i^a \) and chooses \( \sigma_i = a \);\(^3\)
- \( j \) chooses \( \sigma_j = a \) if \( m_i^a \in M_i \cap M_j \), and chooses uniformly from \( A \) if \( m_i^a \in M_i \setminus M_j \).

In this SPE, the payoff of both groups is positive whenever \( a \) is such that \( m_i^a \in M_i \cap M_j \). This event occurs with probability \( |M_i \cap M_j| \), since \( |A| = 1 \). Therefore, the expected payoff of \( i \) is

\[
|M_i \cap M_j|.
\]

In fact, it is easy to verify that all SPEs (except a subset of measure zero of them) have the same derived expected payoffs. If we interpret \( M_i = L_i \cup E_{ij} \), then (1.2) is derived.

### A.1.3 Competitive Interaction

Suppose \( \tau = \text{comp} \), and \( i \) is the sender. The following is an SPE of \( \Gamma_t \):

- \( i \) engages, sends \( m_i^a \), and chooses \( \sigma_i = a \) if and only if \( a \) is such that \( m_i^a \in M_i \setminus M_j \);
- \( j \) chooses \( \sigma_j = a \) if \( m_i^a \in M_i \cap M_j \), and chooses uniformly from \( A \) if \( m_i^a \in M_i \setminus M_j \).

In this SPE, the sender guarantees a non-negative payoff. The payoff is positive whenever \( a \) is such that \( m_i^a \in M_i \setminus M_j \). As a result, the *ex ante* (before the identity of the sender is determined) expected payoff of group \( i \) is

\[
\frac{1}{2} \left[ \Pr \left( \{a : m_i^a \in M_i \setminus M_j\} \right) (2) + \Pr \left( \{a : m_i^a \in M_j \setminus M_i\} \right) (-2) \right] = |M_i \setminus M_j| - |M_j \setminus M_i|.
\]

Again, all SPEs (except a subset of measure zero of them) have the same derived expected payoffs. If we interpret \( M_k = L_k \cup E_k \cup N_k \), \( k = i, j \), then (1.3) is derived.

### A.2 Proofs

#### A.2.1 Proof of Lemma 1.2

*Proof.* Let \( \{L_i^t\} \), be symmetric and localized. We want to show that \( \{L_i^{t+1}\} \), is also symmetric and localized, namely, it satisfies conditions (1.15), (1.16), and (1.11).

It is obvious that \( \{L_i^{t+1}\} \), satisfies condition (1.15) by Proposition 1.1 and Corollary 1.2.

Next, to see that \( \{L_i^{t+1}\} \), satisfies condition (1.16), consider \( i, j, k \) such that \( d_{ij} = d_{ik} \). If \( d_{ij} = d_{ik} \leq d^{*} \), where \( d^{*} \) is determined by (1.14) for every \( t \geq 1 \), then we know by Corollary 1.1 that \( L_i^t \cup E_i^t = L_j^t \cup E_j^t = L_k^t \cup E_k^t \). From the definition of \( L_i^{t+1} \), it follows that \( |L_i^{t+1} \cap L_j^{t+1}| = |L_i^{t+1} \cap L_k^{t+1}| \). Suppose \( d_{ij} = d_{ik} > d^{*} \). There are two cases to consider. First,

\(^3\)Note that this is a weakly dominant strategy for \( i \).
If \( d_{ij} = d_{ik} + d^* + 1 \), then \( i \) will learn nothing from either \( j \) or \( k \) (nor would the latter two learn anything from \( i \)). So \( |L_{ij}^{t+1} \cap L_{jk}^{t+1}| = |L_{ij}^{t+1} \cap L_{ik}^{t+1}| = |L_{ij}^{t} \cap L_{jk}^{t}| = |L_{ij}^{t} \cap L_{ik}^{t}| \). Second, suppose \( d_{ij} = d_{ik} = d^* + 1 \). Observe that

\[
|L_{ij}^{t+1} \cap L_{jk}^{t+1}| = |L_{ij}^{t} \cap L_{jk}^{t}| + |E_{ij}^{t}| + |E_{jk}^{t}| + |E_{ik}^{t} \cap E_{jk}^{t}|
\]

\[
|L_{ij}^{t+1} \cap L_{ik}^{t+1}| = |L_{ij}^{t} \cap L_{ik}^{t}| + |E_{ij}^{t}| + |E_{ik}^{t} \cap E_{jk}^{t}|
\]

where \( E_{m-n} = E_{m} \setminus E_{mn} \). By Assumption 1.1, \( |E_{ij}^{t}| = |E_{ik}^{t}| \). Since the decision problem is symmetric, it must be that \( |E_{ij}^{t}| = |E_{ik}^{t}| \) as well. Moreover, that both \( j \) and \( k \) are \( i \)'s degree \( d^* + 1 \) neighbors implies that \( j \) and \( k \) are within the same \( d^* \) neighborhood. By Assumption 1.2, all groups within a \( d^* \) neighborhood learn the same subset of elements from their degree \( d^* + 1 \) neighbors. It follows that \( E_{i-j}^{t} = E_{k-j}^{t} \) up to a subset of measure zero. Therefore, \( |E_{i-j}^{t} \cap E_{j-i}^{t}| = |E_{i-k}^{t} \cap E_{k-j}^{t}|. \footnote{We do not actually need Assumption 1.2 to prove that \( |E_{i-j}^{t} \cap E_{j-i}^{t}| = |E_{i-k}^{t} \cap E_{k-j}^{t}| \); Assumption 1.1 alone is enough, although the exposition will be more complicated.}

As a result, condition (1.16) holds in \( t + 1 \).

Lastly, condition (1.11) says that if an element \( \ell \) is learnable by \( i \) from a degree \( d \) neighbor, then it cannot be in the dialect of a neighbor of degree greater than \( d \). Suppose, for contradiction, that \( \{L_{ij}^{t+1}\}_i \) violates condition (1.11), i.e. there exists an element \( \hat{\ell} \) such that

\[
\hat{\ell} \in L_{ij}^{t+1} \setminus L_{ij}^{t+1} \quad \text{and} \quad \hat{\ell} \in L_{ik}^{t+1}
\]

for some \( i, j, k \) such that \( d_{ij} < d_{ik} \). Since \( \{L_{ij}^{t}\}_i \) is localized and \( L_{ik}^{t+1} = L_{ij}^{t} \cup E_{ij}^{t} \cap N_{i}^{t} \), it must be the case that either \( \hat{\ell} \in E_{ik}^{t} \setminus E_{ij}^{t} \) or \( \hat{\ell} \in E_{ik}^{t} \setminus E_{ij}^{t} \). \( \hat{\ell} \in E_{ik}^{t} \setminus E_{ij}^{t} \) contradicts the assumption that when \( d_{ik} = d_{jk} \), \( i \) and \( j \) learn the same set from \( k \).

If \( \hat{\ell} \in E_{ik}^{t} \setminus E_{ij}^{t} \) and \( \hat{\ell} \in L_{ik}^{t+1} \setminus L_{ij}^{t+1} \) imply that

\[
\hat{\ell} \in (E_{ik}^{t} \cap (L_{ij}^{t} \cup E_{ij}^{t})) \quad \text{and} \quad \hat{\ell} \notin (E_{ij}^{t} \cap (L_{ij}^{t} \cup E_{ij}^{t}))).
\]

But this is inconsistent with Lemma 1.1 and Proposition 1.2. If \( d_{ij} \leq d^* \), then there is no element in \( L_{ij}^{t} \cup E_{ij}^{t} \) that \( i \) does not know, hence contradicting the second conjunct. If \( d_{ij} > d^* \), \( k \) will not choose to learn anything from \( j \). This is true because when \( d_{ij} < d_{ik} \), we have \( d_{ki} = d_{ij} \). By \( d_{ij} < d_{ik} \) and \( d_{ij} \leq d^* \), it must be the case that \( d_{ik} > d^* + 1 \). Combining this fact with Lemma 1.1, \( k \) will not learn anything from \( i \) and so neither will \( k \) learn from \( j \). Hence we have a contradiction with the first conjunct. Therefore, condition (1.11) must hold for \( \{L_{ij}^{t+1}\}_i \).

This completes the proof. \( \square \)

### A.2.2 Proof of Proposition 1.3

To prove this proposition, it is helpful to first introduce the following lemma:

**Lemma A.1.** Let \( \{L_{ij}^{t}\}_{i \in \mathcal{I}} \) be symmetric and localized. Then, \( d^* \leq d_{ij} \), where \( d^* \) is determined by (1.14) for every \( t \geq 1 \).

To prove this proposition, it is helpful to first introduce the following lemma:
Proof. First, observe that, for all \( t \geq 1 \), all \( i \in \mathcal{G} \),
\[
\sum_{k=1}^{d} |P_i^j(k)| \geq (2^d - 1)|N^*|, \quad \forall d \in \{1, \ldots, D\}.
\] (A.3)

This is true because, from Proposition 1.1, we know that each group invents a measure of \(|N^*|\) new linguistic elements in every \( t - 1 \) (\( t = 1, 2, \ldots \)). Hence in \( t \), the set of learnable elements within a degree \( d \) neighborhood for group \( i \), \( \bigcup_{k=1}^{d} P_i^j(k) \), must contain at least, and potentially more than, a measure of \((2^d - 1)|N^*|\) elements. Recall that \( d^* \) and \( \bar{d} \) are determined by conditions (1.14) and (1.19), respectively. According to (1.19), \(|P_i^j(d)| = 2^{d-1}|N^*|\) for \( d \leq \bar{d} \), and so (A.3) holds with equality for \( d \leq \bar{d} \). Therefore, conditions (1.14) and (1.19) imply that \( d^* \leq \bar{d} \) for all \( t \geq 1 \). \( \square \)

This lemma says that regardless of initial linguistic composition, no group will ever learn the all the learnable elements of all of its \( \bar{d} + 1 \) degree neighbors.

Now we are ready to prove Proposition 1.3 itself.

Proof. First, consider \( i, j \) such that \( d_{ij} > \bar{d} + 1 \). From Lemma A.1, \( d^* \leq \bar{d} \). Trivially, \( d^* + 1 \leq \bar{d} + 1 \). Therefore, if \( d_{ij} > \bar{d} + 1 \), then \( d_{ij} > d^* + 1 \). According to Corollary 1.1, \( i \) will never learn any elements of \( j \)'s language. Every period, \( |L_j \setminus L_i| \) increases by exactly \( |N^*| \).

Next, consider \( i, j \) such that \( d_{ij} = \bar{d} + 1 \) and \( d^* < \bar{d} \). Since \( d^* + 1 < d_{ij} \), by Corollary 1.1 \( i \) will not learn any element of \( j \)'s language.

Next consider \( i, j \) such that \( d_{ij} = \bar{d} + 1 \) and \( d^* = \bar{d} \). From 1.9, \( |P_i(d)| \geq 2^{d-1}|N^*| \) for all \( d \). From Corollary 1.1, \( i \) learns \( |E_i^*| - \sum_{k=0}^{d^*-1} |P_i^j(k)| \) from its \( d^* + 1 \) neighbors. The amount that \( i \) learns from its degree \( \bar{d} + 1 \) neighbors, \( |E_i^*| - \sum_{k=0}^{d^*} |P_i^j(k)| \), is maximized when \( |P_i(d)| = 2^{d-1}|N^*| \) for all \( d \leq \bar{d} \). Since \( d^* = \bar{d} \) it must be true by Proposition 1.2 that \( |E_i^*| < \sum_{k=0}^{d^*+1} 2^{k-1}|N^*| \). According to Assumption 1.1, \( i \) learns an equal measure of each of its degree \( \bar{d} + 1 \) neighbor’s languages, so \( i \) will not learn more than \(|N^*|\) of any single one of its degree \( \bar{d} + 1 \) neighbors’ languages. Therefore \( |E_i^*| < |N_j| \) and \( |L_j \setminus L_i| \) increases every period. \( \square \)

A.2.3 Proof of Proposition 1.4

Proof. According to Definition 1.6, we need to show that if condition (1.20) holds for \( t \) then it also holds for \( t + 1 \).

Suppose (1.20) is true for \( t \). We know from Proposition 1.1 that between \( t \) and \( t + 1 \), each group invents a set of new elements of measure \(|N^*|\). Therefore, within a degree \( \bar{d} \) neighborhood, a measure of \( 2^\bar{d}|N^*| \) new elements will have been invented by the end of period \( t \). For \( \bar{d} = 0 \), it is trivially true that (1.20) holds for \( t + 1 \), because a dialect is always the same as itself. Consider \( \bar{d} \geq 1 \). At \( t + 1 \), dialects within a \( \bar{d} \) neighborhood will stay as one language—i.e. \( L_i^{t+1} \cup E_i^{t+1} = L_i^{t+1} \cup E_i^{t+1} \) for all \( i, j \) such that \( d_{ij} \leq \bar{d} \)—if and only if the marginal benefit at \( |E_i^*| = (2^\bar{d} - 1)|N^*| \) outweighs the marginal cost. But this is true by how \( \bar{d} \) is determined in (1.19). Therefore, condition (1.20) holds for \( t + 1 \). \( \square \)

A.2.4 Proof of Proposition 1.5

One additional lemma will be useful in proving Proposition 1.5.
Lemma A.2. For any $t \geq 1$ and $\overline{d} \geq 1$, we have
\[
d^* < \overline{d} \Rightarrow |E_t^*| > (2^{\overline{d}} - 1)|N^*|,
\]
and
\[
d^* = \overline{d} \Rightarrow |E_t^*| \geq (2^{\overline{d}} - 1)|N^*|.
\]

Proof. Let $|\overline{E}|$ denote the measure of elements learned by a group in the steady state $\overline{X}$. From Proposition 1.4, we know that $|\overline{E}|$ is at least $(2^{\overline{d}} - 1)|N^*|$. Consider the first implication. Suppose $d^* < \overline{d}$, so that the region’s linguistic composition is out of the steady state. Then the marginal benefit of learning elements on the interval $[(2^{\overline{d}} - 1)|N^*|, |\overline{E}|]$ is at least $\frac{r + \beta}{1 + \gamma} \pi(d^*)$, which is strictly higher than the marginal benefit of the elements on the same interval if a group were learning in the steady state. This is illustrated in Figure A.1. In the top panel, the linguistic composition if out of steady state ($d^* = 2$); in the bottom panel, the linguistic composition is in steady state ($\overline{d} = 3$). On the interval $[(2^{\overline{d}} - 1)|N^*|, |\overline{E}|]$, the marginal benefit in the top panel is higher than that in the bottom panel. Since the marginal cost of learning is the same both in and out of the steady state, it follows that a group must learn a strictly larger measure of elements out of steady state. As a consequence, $|E_t^*| > |\overline{E}| \geq (2^{\overline{d}} - 1)|N^*|$ and the first implication is established. The second implication follows trivially from the fact that when $d^* = \overline{d}$, the lower bound in Proposition 1.2 applies to both $|E_t^*|$ and $|\overline{E}|$. But since the latter is bounded below by $(2^{\overline{d}} - 1)|N^*|$ according to Proposition 1.4, so must $|E_t^*|$.

Now we are ready to prove Proposition 1.5 itself.

Proof of Proposition 1.5. If $\overline{d} = 0$, then by (1.14) and (1.19) we have $d^* = 0$ for all $t \geq 1$. Consequently, the steady state $\overline{X}$ is achieved at the end of the first period at the latest. Suppose $\{L_i^0\}_i$ is such that $L_i^0 = L_j^0$ for all $i, j \in \mathcal{G} \subseteq \mathcal{G}$. Then, at the beginning of $t = 1$, each $L_k$ will differ by at least a measure of $|N^*|$ elements. Since $d^* = 0$ for all $t$, we must have $|E_t^*| < |N^*|$. Thus, for all $t \geq 2$, it must be the case that $L_i^1 \neq L_j^1$ for all $i, j \in \mathcal{G}$. If, on the other hand, not all of the initial set of dialects are identical and $d^* = 0$, then a fortiori, $L_i^1 \neq L_j^1$ for any $i, j \in \mathcal{G}$ and any $t \geq 1$.

For the remainder of the proof, we assume $\overline{d} \geq 1$, and the proof proceeds as follows: First we show that after the initial period, all pairs of groups $i$ and $j$ with $d_{ij} > \overline{d}$ will never speak the same language. Then, we show that all groups within a degree $\overline{d}$ neighborhood will eventually share a common language as defined in Definition 1.2.

Let $\{L_i^t\}_i$ be symmetric and localized, and consider a pair of groups $i, j$ with $d_{ij} > \overline{d}$. By Lemma A.1, we know that $d_{ij} > d^*$. This implies that, for each $t$, the measure of $i$’s set of learnable elements is strictly greater than the measure of the set that $i$ actually learns: $\sum_{k=0}^{d_{ij}} |p_i^t(k)| > |E_t^i|$. According to Lemma 1.1, therefore, it is never optimal for $i$ to learn the entire set of $L_j^i \setminus L_i^i$ for every $t \geq 0$. As a consequence, we have
\[
d_{ij} > \overline{d} \Rightarrow L_i^t \cup E_i^t \neq L_j^t \cup E_j^t, \quad \forall t \geq 1.
\]

---

To simplify the drawing, we assumed that the marginal benefit of learning has only one step for each $d_{ij}$. See footnote 36 for more detail.
$\lambda^*$

Out of steady state $\bar{\lambda}$

In steady state $\bar{\lambda}$

Figure A.1: Marginal Benefit of Learning in and out of Steady State
Next, consider a pair of groups $i, j$ for which $d_{ij} \leq \bar{d}$. In each period $t$, $i$’s closest $2^{\bar{d}} - 1$ neighbors invent a measure of $(2^{\bar{d}} - 1)|N^*|$ new linguistic elements. Lemma A.2 shows that when $d_t^i$ is less than (or equal to) $\bar{d}$, the size of the set of elements learned by $i$ is strictly (or weakly) greater than the number of elements newly invented by the closest $2^{\bar{d}} - 1$ neighbors. Since each $L_i$ is of a finite measure, and the rate of linguistic convergence (i.e. size of acquired elements from existing dialects) is higher than the rate of linguistic divergence (i.e. size of invented elements) within a degree $\bar{d}$ neighborhood, dialects within such a neighborhood will attain $L_i \cup E_i = L_j \cup E_j$ in finitely many periods and will remain that way thereafter. As a result, there exists a $T < \infty$ such that for all $t \geq T$, we have $L_i^t \cup E_i^t = L_j^t \cup E_j^t$ for all $i, j$ with $d_{ij} \leq \bar{d}$. This completes the proof.

\[\square\]

A.3 Figures
Figure A.2: Distribution of Languages in the World
(a) Divergence of Malayo-Polynesian Languages

(b) Divergence of Indo-European Languages

Figure A.3: Estimated Linguistic Divergence in Malayo-Polynesian and Indo-European Language Families
Each dot is a pair of groups. Note that the vertical axis is in log-scale. On both graphs, the rate of divergence is very high immediately after a split: pairs of languages seem to explode apart, and later settle down to a much low rate of divergence. Source: Pagel (2000, p.399–400).
Appendix B

Supplementary Materials for Chapter 2

B.1 Experiment Instructions
Instructions

Welcome! This is an experiment in the economics of decision making. Please read the following instructions carefully. A good understanding of these instructions and well reasoned decisions during the experiment will earn you a considerable amount of CASH at the end of the experiment.

When you finish reading the instructions, there will be a quiz to test your understanding. Your performance in this quiz will partly determine your cash earnings at the end.

If you have any question about these instructions, please raise your hand. The experimenter will come to you and answer your question privately.

Your Role

Throughout this experiment, you are known as Person RED, and you will be interacting with another participant in this room, known as Person BLUE.

Timeline

This experiment consists of a number of similar rounds, and each round has two (2) Phases.

At the beginning of Phase 1 of each round, each person is given a set of messages that refers to the actions (A, B or C), which they can take in an interaction in Phase 2. In Phase 1 of each round you have the opportunity to make changes to this set of messages. The specifics will be explained in detail later.

In Phase 2 of each round you and Person BLUE will participate in an interaction, and each of you has to choose an action to generate payoffs. Payoffs from the interaction are determined by the combination of your action and Person BLUE’s action.

Figure 1 on page 2 illustrates how the experiment will proceed.

To help you best understand this process, we first go through how Phase 2 works, and then explain the procedure in Phase 1 afterwards.
Figure 1: Timeline of Experiment

**Phase 2: Two-Person Interaction**

The interaction in Phase 2 is represented as a $3 \times 3$ table similar to the one below, where **Person RED**’s actions are represented using rows and **Person BLUE**’s actions are represented using columns:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>−15 ED</td>
<td>15 ED</td>
<td>15 ED</td>
</tr>
<tr>
<td>B</td>
<td>15 ED</td>
<td>−15 ED</td>
<td>15 ED</td>
</tr>
<tr>
<td>C</td>
<td>−15 ED</td>
<td>15 ED</td>
<td>−15 ED</td>
</tr>
</tbody>
</table>

In other words, **Person RED** chooses among **rows** and **Person BLUE** chooses among **columns**. The interaction payoffs, in **experimental dollars** or **ED**, are indicated in the table above by the cells with numbers. The first number in each cell (with color red) denotes the payoff to **Person RED**, and the second number in each cell (with color blue) denotes the payoff to **Person BLUE**. This means:

<table>
<thead>
<tr>
<th>Choices</th>
<th>Payoffs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Person RED</strong> chooses</td>
<td><strong>Person BLUE</strong> chooses</td>
</tr>
<tr>
<td>A (Row 1)</td>
<td>A (Column 1)</td>
</tr>
<tr>
<td>A (Row 1)</td>
<td>B (Column 2)</td>
</tr>
<tr>
<td>A (Row 1)</td>
<td>C (Column 3)</td>
</tr>
<tr>
<td>B (Row 2)</td>
<td>A (Column 1)</td>
</tr>
<tr>
<td>B (Row 2)</td>
<td>B (Column 2)</td>
</tr>
<tr>
<td>B (Row 2)</td>
<td>C (Column 3)</td>
</tr>
<tr>
<td>C (Row 3)</td>
<td>A (Column 1)</td>
</tr>
<tr>
<td>C (Row 3)</td>
<td>B (Column 2)</td>
</tr>
<tr>
<td>C (Row 3)</td>
<td>C (Column 3)</td>
</tr>
</tbody>
</table>

A negative payoff means losing ED. The payoff table will be the same in every round.
Now please turn to the computer screen. What you’re seeing is the typical screen for Phase 1. At the top of the screen you can see which round you’re in, your starting payoff, and your role. The second line indicates what you need to do at this stage. We will come back and explain the other layout features of this screen later. For now please click on the “SUBMIT” button at the bottom to skip to Phase 2.

You’re now seeing the decision screen for Phase 2. On the right side of the screen is the payoff table we explained earlier. On the left side, there are two boxes called PUBLIC LIST and PRIVATE LIST. The symbols in each box are your messages, and the symbols under the heading “A” refer to action A in the payoff table, and similarly for symbols under headings “B” and “C”.

In this phase you need to decide which action to take. Your decision is implemented in the following manner:

i) Both of you submit an initial choice of action to the computer by clicking on a message that refers to that action. The difference between the public and private lists is that messages in the public list can be understood by both you and Person BLUE, while messages in the private list can be understood only by you and not by Person BLUE. In other words, if you want Person BLUE to know what action you intend to take, you should send a message from the public list; and if you do not want Person BLUE to know your intended action, you should send a message from your private list.

ii) The computer then flips a coin, and decides which one of you will receive a revision opportunity. If you get this opportunity, you will be able to observe the message sent by the other person, and then have the chance to revise your initial choice based on the other person’s submitted message.

iii) The person who receives this revision opportunity would submit his/her final choice. (The other person’s initial choice is final.)

iv) The interaction payoff is determined based on the final choices.

Let’s go over an example. Suppose you want to choose action B, and the only available message for B is in the public list. Then you would click on the symbol under the heading “B” in the public list, and then hit the “SUBMIT” button.

After both you and Person BLUE have submitted a choice, the computer will flip a coin. If the coin comes up heads, Person RED will get the revision opportunity; if the coin comes up tails, Person BLUE will get the revision opportunity. The person who doesn’t get the revision opportunity will need to wait until the other person submits his/her final choice.

Suppose you get the revision opportunity. You would see the message sent by the other person to the computer. If the other person’s message was sent from his/her public list, the screen will
If the other person’s message was sent from his/her private list, the screen will display:

<table>
<thead>
<tr>
<th>The other person sent the following message to the computer:</th>
</tr>
</thead>
<tbody>
<tr>
<td>MESSAGE:  أنحاء ACTION INTENDED: UNKNOWN</td>
</tr>
</tbody>
</table>

Based on this information, you need to make a final decision on which action to take. To complete this example, for those who have received the revision opportunity, please choose an action (that you think would maximize your payoff) by clicking on a message and then the “SUBMIT” button. You may stick to your initial choice or choose a different action as you see fit.

After the revision is finished, the outcome of the interaction will be announced, and your payoff for this round is displayed at the bottom right corner of the screen.

Next, we will go over how Phase 1 works. Please click on the “NEXT ROUND” button at the lower right corner.

**Phase 1: Changing Your Message List**

What you’re seeing now is the Phase 1 decision screen for a new round. In each new round, you will be given a set of messages (i.e. symbols) that refers to actions in the Phase 2 interaction with Person BLUE, which we went over earlier.

Towards the middle of the screen, you can see two boxes under the names “PUBLIC LIST” and “PRIVATE LIST”. As explained earlier, the symbols under the heading “A” refer to action A in the Phase 2 interaction, and similarly for symbols under headings “B” and “C”. The difference between the public and private lists is that messages in the public list can be understood by both you and Person BLUE, while messages in the private list can be understood only by you and not by Person BLUE.

Person BLUE’s messages are similarly divided into these two lists. Moreover, Person BLUE’s public list consists of exactly the same messages as your public list.

The set of messages given to you will be different in every round. This means sometimes you may have several messages referring to the same action, and sometimes no message that refers to a particular action. Also, the same symbol may refer to different actions in different rounds.
In Phase 1, you have **two decisions** to make regarding your messages.

(i) **Learning messages from Person BLUE.** This is done by choosing a number in the first selection box. The cost of learning, indicated as “(LEARNING COST: #)”, will be automatically calculated as you choose the number. The more you learn, the higher the cost to you will be.

You can learn as many messages as there are in Person BLUE’s private list. The number of messages in Person BLUE’s private list is shown in the first line as “(AVAILABLE: #)”.

If you chose to learn \(N\) messages from Person BLUE, then \(N\) messages from Person BLUE’s private list will be randomly selected and moved to the public list. In other words, learning means to move messages from Person BLUE’s private list (which you could not see) to the public list (which both of you can see). At the beginning of Phase 2, you will know if the other person has learned from you in Phase 1 by observing the changes in the two lists.

For practice, let’s all choose to learn one message by selecting “1” in the first selection box.

(ii) **Creating messages for your private list.** This is done by choosing numbers in the bottom three selection boxes on the screen. You can create at most three new messages in each round, and you can choose which action you want the new message(s) to refer to by selecting numbers in the appropriate selection boxes.

The cost of message creation, indicated as “(CREATING COST: #)”, is automatically calculated for you as you choose the number. The more messages you create, the higher the cost.

For practice, let’s suppose you want to create one message for action B. Then you would select “1” in the second last selection box, and leave the other selection boxes corresponding to actions A and C at “0”.

To summarize, in Phase 1 you can either learn or create messages. Learning moves messages from Person BLUE’s private list to the public list (so that you may understand them), while message creation is to generate new messages in your private list (so that Person BLUE may not understand them).

After you finish both decisions in this phase, you can click the “SUBMIT” button. Then both the public and private lists will be updated according to your choices and the choices of Person BLUE.

Now click the “SUBMIT” button to enter Phase 2, and use this opportunity to familiarize yourself with the Phase 2 procedure.

Before you take the quiz, you will have two more practice rounds to help you better understand procedure.
Round Payoff

You will begin each round with a starting payoff of 30 ED. This amount is the same in each round. Your round final payoff will be calculated according to the following formula:

\[ 30 - \text{cost of learning/creating messages in Phase 1 + interaction payoff in Phase 2} \]

Your Cash Earnings

At the end of the experiment, you will see a table summarizing your quiz result and the final payoffs in each round. One of these payoffs will be randomly selected to determine your cash earnings. The random draw works as follows: When the experiment finishes, the experimenter will come to your desk and roll a die to determine which round you will get paid for. For example, if the die turns up “1”, then your final payoff in Round 1 will be converted to Canadian dollars and paid to you.

Your total cash earning at the end of the experiment will consist of following three items:

- $7 show-up fee
- Final payoff of one randomly selected round
- 1 ED \times \text{number of correctly answered quiz questions}

The last two items are measured in ED, which will be converted to Canadian dollars (CAD) at the rate of 1 ED = 0.35 CAD. The amount in Canadian dollars will be paid to you in cash at the end of this experiment.

Remember, each round could be the one determining your final cash earnings. SO TREAT THE DECISIONS IN EACH ROUND AS IF THEY WERE THE ONES DETERMINING YOUR FINAL CASH EARNINGS.
Summary

- In this experiment you are Person RED, and you will be interacting with another participant in this room, known as Person BLUE.

- The experiment consists of a number of similar rounds, and each round has two Phases.

- Phase 1 involves changing a set of messages by either creating new ones in your private list or learning them from Person BLUE’s private list (thereby moving messages in Person BLUE’s private list to the public list).

- Learning and creating messages are both costly.

- Messages in the public list are the same for both persons, so you both know what actions they refer to. However, neither of you can tell which actions the messages in the other person’s private list refer to.

- You and Person BLUE will be given new sets of messages in each round.

- Phase 2 involves an interaction with Person BLUE. You need to choose an action by clicking a message corresponding to the desired action in either public or private list.

- In the table summarizing the actions and payoffs of the interaction, the rows represent Person RED’s actions, and the columns represent Person BLUE’s actions. The first number in each cell (in red) indicates payoff to Person RED and the second number (in blue) indicates payoff to Person BLUE.

- After initially submitting a message to the computer, there is a 50% chance that you get an opportunity to see what message was sent by the other person, and then revise your initial action accordingly; but there is also a 50% chance that the other person gets the revision opportunity and changes his/her action based on the message you sent to the computer.

- In any given round, only one of you will get the revision opportunity.

- The final payoff in each round will be determined according to

\[ 30 - \text{cost of learning/creating messages in Phase 1} + \text{interaction payoff in Phase 2}, \]

- At the end of the experiment, the final payoff in one randomly selected round will be used to determine your cash earnings.

- Suppose your final payoff in the randomly selected round is \( X \), and you correctly answered \( N \) questions in the quiz, your total cash earning in Canadian dollars will be

\[ 7 + 0.35(X + N) \] Canadian dollars.
Appendix C

Supplementary Materials for Chapter 3

C.1 Proof of Lemma 3.1

Proof. We first rewrite the inner maximization of (3.4) as

$$\min_{a \in \mathbb{R}} \int_{\mathbb{R}} (a - \mu)^2 f(\mu \mid \beta_m, \beta_s) d\mu,$$

which can then be simplified into an expression that involves only the posterior mean and variance

$$\min_{a \in \mathbb{R}} \left[ (a - \beta_m)^2 + \beta_s \right].$$  \(\text{(C.1)}\)

It is clear from (C.1) that the optimal solution to set \(a\) equal to the posterior mean:

$$a^A(n^A) = \beta_m = \frac{\mu' \eta' + n^A \bar{x}}{\eta' + n^A}.$$

Substituting \(a^A(n^A)\) into the original problem in (3.4), we have

$$\max_{n \in \mathbb{R}^+} \mathbb{E}_{\bar{x}} \left[ \int_{\mathbb{R}} Q(\mu)f(\mu \mid \beta_m, \beta_s) d\mu - \beta_s - nc \right].$$  \(\text{(C.2)}\)

Since the last two terms in the square bracket in (C.2) do not depend on \(\bar{x}\), and that for every \(\mu \in \mathbb{R}, \mathbb{E}_{\bar{x}} f(\mu \mid \beta_m, \beta_s) = f(\mu \mid \mu', s)\), we derive

$$\max_{n \in \mathbb{R}^+} Q^* - \beta_s - nc, \quad \text{or} \quad \max_{n \in \mathbb{R}^+} Q^* - \frac{s \eta'}{\eta' + n} - nc,$$  \(\text{(C.3)}\)

where \(Q^* = \int_{\mathbb{R}} Q(\mu)f(\mu \mid \mu', s) d\mu\), a quantity that does not depend on \(n\). Hence it is straightforward to verify from the first order condition

$$\frac{\nu'}{(\eta' + n)^2} - c = 0$$
that
\[ n^A = \left[ \frac{\nu'}{c} \right]^{\frac{1}{2}} - \eta' . \]

Since (C.1) is strictly convex in \( a \) and (C.2) is strictly concave in \( n \), the derived solutions indeed maximize (3.4). Lemma 3.2 can be proved similarly.

### C.2 Proof of Lemma 3.3

**Proof.** The logic is the same as the preceding proof. In particular, \( a^A(n^A) \) is derived in the same way as \( a^A(n^A) \). The derivation of \( n^A \) is slightly different, due to the presence of the utility for surprise. The first order condition characterizing \( n^A \) is

\[
\sigma \frac{\partial}{\partial n} \mathbb{E}_x [U_{\text{surp}}(n; \bar{x})] + \frac{\nu'}{(\eta' + n)^2} - c = 0. \tag{C.4}
\]

Since both \( f(\cdot \mid \mu', s) \) and \( f(\cdot \mid \beta_m, \beta_s) \) are Gaussian, it follows that

\[
U_{\text{surp}}(n; \bar{x}) = \int_{\mathbb{R}} \ln \left( \frac{f(\mu \mid \beta_m, \beta_s)}{f(\mu \mid \mu', s)} \right) f(\mu \mid \beta_m, \beta_s) d\mu
\]

\[
= \int_{\mathbb{R}} \ln f(\mu \mid \beta_m, \beta_s) f(\mu \mid \beta_m, \beta_s) d\mu - \int_{\mathbb{R}} \ln f(\mu \mid \mu', s) f(\mu \mid \beta_m, \beta_s) d\mu
\]

\[
= \ln \left( \frac{1}{\sqrt{2\pi \beta_s}} \right) - \int_{\mathbb{R}} \left( \frac{\mu - \beta_m}{2 \beta_s} \right)^2 f(\mu \mid \beta_m, \beta_s) d\mu
\]

\[
- \ln \left( \frac{1}{\sqrt{2\pi s}} \right) + \int_{\mathbb{R}} \left( \frac{\mu - \mu'}{2 s} \right)^2 f(\mu \mid \beta_m, \beta_s) d\mu
\]

\[
= \frac{1}{2} \ln s - \frac{1}{2s} \int_{\mathbb{R}} (\mu^2 - \beta_m^2 + \beta_m^2 - 2\mu\mu' + (\mu')^2) f(\mu \mid \beta_m, \beta_s) d\mu
\]

\[
= \frac{1}{2} \ln s - \frac{1}{2s} \left[ \beta_s + (\beta_m - \mu')^2 \right]
\]

\[
= \frac{1}{2} \left[ \ln s - \ln \nu' + \ln(\eta' + n) + \frac{\eta'}{\eta' + n} + \frac{1}{s} \left( \frac{\eta' \eta' + n n}{\eta' + n} - \mu' \right)^2 - 1 \right].
\]

**Pratt et al. (1995)** show that, when the prior is normal with mean \( \mu' \) and variance \( s \) and the signal distribution has variance \( \nu' \), the sample mean of \( n \) i.i.d. signals is normally
distributed with mean $\mu'$ and variance $\zeta' = \frac{1}{n}(ns + \nu')$. Thus,

$$
E_x[U_{\text{surp}}(n; \bar{x})] = \frac{1}{2} \left[ \ln s - \ln \nu' + \ln(\eta' + n) + \frac{\eta'}{\eta' + n} \\
+ \frac{1}{s} \int_R \left( \frac{\mu' \eta' + n \bar{x}}{\eta' + n} - \mu' \right)^2 f(\bar{x} | \mu', \zeta') \, d\bar{x} - 1 \right].
$$

Observe that

$$
\frac{1}{s} \int_R \left( \frac{\mu' \eta' + n \bar{x}}{\eta' + n} - \mu' \right)^2 f(\bar{x} | \mu', \zeta') \, d\bar{x} = \frac{1}{s} \left( \frac{n}{\eta' + n} \right)^2 \int_R (\bar{x} - \mu')^2 f(\bar{x} | \mu', \zeta') \, d\bar{x} = \frac{1}{s} \left( \frac{n}{\eta' + n} \right)^2 \left( \frac{ns + \nu'}{n} \right) = \left( \frac{n}{\eta' + n} \right)^2 \left( 1 + \frac{\eta'}{n} \right).
$$

Thus,

$$
E_x[U_{\text{surp}}(n; \bar{x})] = \frac{1}{2} \left[ \ln s - \ln \nu' + \ln(\eta' + n) + \frac{\eta'}{\eta' + n} + \left( \frac{n}{\eta' + n} \right)^2 \left( 1 + \frac{\eta'}{n} \right) - 1 \right]. \tag{C.5}
$$

Evaluating the first summand in (C.4) using (C.5), we get

$$
\frac{\partial}{\partial n} E_x[U_{\text{surp}}(n; \bar{x})] = \frac{1}{2} \left[ \frac{1}{\eta' + n} - \frac{\eta'}{\eta' + n}^2 \\
+ 2 \left( \frac{n}{\eta' + n} \right) \left( \frac{\eta'}{\eta' + n}^2 \right) \left( 1 + \frac{\eta'}{n} \right) - \left( \frac{n}{\eta' + n} \right)^2 \frac{\eta'}{n^2} \right] \\
= \frac{1}{2} \left[ \frac{n}{(\eta' + n)^2} + 2n \eta' - \frac{\eta'}{\eta' + n} \right] = \frac{1}{2(\eta' + n)}. \tag{C.6}
$$

Substituting (C.6) back into (C.4) simplifies the first order condition to

$$
\frac{\sigma^2}{2} (\eta' + n) + s \eta' - c(\eta' + n)^2 \bigg|_{n = n^{A^*}} = 0.
$$

Solving for $n^{A^*}$ yields the expression in (3.12). From (C.4) and (C.6), we can also easily verify that the second order condition is satisfied. Therefore, $n^{A^*}(\sigma)$ indeed maximizes (3.11).
C.3 First and Second Order Conditions for (3.14)

Applying some algebraic manipulations to (3.14), we get

\[
\min_n \Pi(n) = \left(\frac{\eta' - \eta}{\eta' + n}\right)^2 \left(\frac{sn}{\eta + n}\right) + \frac{sn}{\eta + n} + nc.
\]

Differentiating \(\Pi(n)\) with respect to \(n\), we get

\[
\frac{\partial}{\partial n} \Pi(n) = -2 \frac{(\eta' - \eta)^2}{(\eta' + n)^3} \frac{sn}{\eta + n} + \frac{(\eta' - \eta)^2}{(\eta' + n)^2} \frac{sn}{(\eta + n)^2} - \frac{sn}{(\eta + n)^2} + c
\]

\[
= \frac{s(\eta' - \eta)^2[-2n(\eta + n) + \eta(\eta' + n)]}{(\eta' + n)^3(\eta + n)^2} - \frac{sn}{(\eta + n)^2} + c.
\]  
(C.7)

Differentiate (C.7) with respect to \(n\) again to get the second order condition:

\[
\frac{\partial^2}{\partial n^2} \Pi(n) = 2 \frac{(3\eta'^2 - 2\eta' + \eta)n}{(\eta' + n)^4},
\]

which is always positive for \(n \geq 0\). Therefore, the first order condition characterizes the unique solution to (3.14). The following quintic equation characterizes the solution \(\hat{n}\):

\[
s(\eta' - \eta)^2(\eta\eta' - \eta\hat{n} - 2\hat{n}^2) - s\eta(\eta' + \hat{n})^3 + c(\eta' + \hat{n})^3(\eta + \hat{n})^2 = 0. \quad (C.8)
\]

C.4 Proof of Theorem 3.1

Proof. Thanks to the strict convexity of \(\Pi(n)\) on \(\mathbb{R}_+\), to show that \(\hat{n} > n^A\), we only need to demonstrate that \(\frac{\partial}{\partial n} \Pi(n)\big|_{n=n^A} < 0\). Evaluating (C.8) at \(n^A\), we have

\[
s(\eta' - \eta)^2\left(\eta\eta' - \eta\left(\sqrt{\eta'c - \eta'}\right) - 2\left(\sqrt{\eta'c} - \eta'\right)^2\right)
- sn\left(\eta' + \sqrt{\eta'c} - \eta'\right)^3 + c\left(\eta' + \sqrt{\eta'c} - \eta'\right)^3 \left(\eta + \sqrt{\eta'c} - \eta'\right)^2,
\]

which simplifies to

\[
(\eta' - \eta)^2 \left[2\eta\eta' - \eta\left(\frac{s}{c}\eta'\right)^{1/2} - 2\left(\left(\frac{s}{c}\eta'\right)^{1/2} - \eta'\right)^2\right]
- \eta\left(\frac{s}{c}\eta'\right)^{3/2} + \left(\frac{s}{c}\eta'\right)^{1/2}\left(\eta' - \eta - \left(\frac{s}{c}\eta'\right)^{1/2}\right)^2. \quad (C.9)
\]

Thus we have arrived at the condition in (3.16). □
C.5 Proof of Proposition 3.1

We take two steps to show that the expression in (3.16) is negative under the stated condition in the proposition. In the first step, we argue that (3.16) is negative whenever \( c = s/\eta' \) and \( \nu' > \nu \). Then, in the second step we show that the derivative of (3.16) with respect to \( c \) is strictly positive under the stated condition. Since \( c = s/\eta' \) implies \( n^A = 0 \), and \( c < s/\eta' \) implies \( n^A > 0 \), the two steps together imply that (3.16) is always negative for \( n^A \geq 0 \).\(^1\) In other words, \( \hat{n} > n^A \) whenever \( \nu' - \nu \geq 3 (\nu'/c)^{1/2} \), the condition under which the derivative of (3.16) with respect to \( c \) is positive.

**Step 1.** At \( c = s/\eta' \), we have \( s/c = \eta' \) and (3.16) simplifies to
\[
\eta^2 \eta' (\eta - \eta') < 0,
\]
because \( \eta' > \eta \) (or \( \nu' > \nu \)), an assumption imposed throughout the paper.

**Step 2.** Differentiating (3.16) with respect to \( c \), we get
\[
(\eta' - \eta) \left[ \frac{\eta}{2c} \left( \frac{s}{c} \eta' \right)^{1/2} + \frac{2}{c} \left( \left( \frac{s}{c} \eta' \right)^{1/2} - \eta' \right) \left( \frac{s}{c} \eta' \right)^{1/2} \right] + \frac{3\eta}{2c} \left( \frac{s}{c} \eta' \right)^{3/2}

+ \left( \frac{s}{c} \right)^{1/2} \left( \eta' \right)^{3/2} \left[ - \left( \eta' - \eta - \left( \frac{s}{c} \eta' \right)^{1/2} \right)^2 + 2 \left( \eta' - \eta - \left( \frac{s}{c} \eta' \right)^{1/2} \right) \left( \frac{s}{c} \eta' \right)^{1/2} \right] \right]. \quad (C.10)
\]
The first line of (C.10) is positive, since all parameters are positive.\(^2\) The second line of (C.10) is non-negative when
\[
\nu' - \nu \geq 3 (\nu'/c)^{1/2}. \quad (C.11)
\]
Therefore, as long as (C.11) holds, \( \frac{\partial}{\partial n} \Pi(n) \bigg|_{n=n^A} < 0 \) for any \( n^A \geq 0 \). This completes the proof.

**Remark C.1.** From (C.10) it is evident that (C.11) is only a sufficient condition for (C.10) to be positive. With some messy algebra, we should be able to derive a more lenient restriction on the gap between \( \nu' \) and \( \nu \).

\(^1\)If \( c \) is too large, \( c \geq \frac{s}{\eta} \left( \frac{2\nu - \eta}{\eta} \right) \) to be exact, then \( \hat{n} = n^A = 0 \), as there would be no benefit to acquiring information. But this is uninteresting. Thus we do not focus on the case where \( c > s/\eta' \).
\(^2\)Recall that \( n^A = \sqrt{s\eta'/c} - \eta' \geq 0 \).