MODELING AND PREDICTING THE EFFECT OF CULTURE IN COMMUNICATION: A MIXED STUDY USING NAMING GAME AND SOCIAL NETWORKS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÖZGE NİLAY YALÇIN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COGNITIVE SCIENCE

SEPTEMBER 2014

Approval of the thesis:

MODELING AND PREDICTING THE EFFECT OF CULTURE IN COMMUNICATION: A MIXED STUDY USING NAMING GAME AND SOCIAL NETWORKS

submitted by **ÖZGE NİLAY YALÇIN** in partial fulfillment of the requirements for the degree of **Master of Science in Cognitive Science Department, Middle East Technical University** by,

Prof. Dr. Nazife Baykal
Dean, Graduate School of **Informatics** ______________

Prof. Dr. Cem Bozşahin
Head of Department, **Cognitive Science** ______________

Prof. Dr. Cem Bozşahin
Supervisor, **Cognitive Science Department, METU** ______________

Examining Committee Members:

Prof. Dr. Deniz Zeyrek
Cognitive Science Department, METU ______________

Prof. Dr. Cem Bozşahin
Cognitive Science Department, METU ______________

Assist. Prof. Dr. Cengiz Acartürk
Cognitive Science Department, METU ______________

Assist. Prof. Dr. Murat Perit Çakır
Cognitive Science Department, METU ______________

Dr. Ayşenur Birtürk
Computer Engineering Department, METU ______________

Date: ______________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: ÖZGE NİLAY YALÇIN

Signature :

ABSTRACT

MODELING AND PREDICTING THE EFFECT OF CULTURE IN COMMUNICATION: A MIXED STUDY USING NAMING GAME AND SOCIAL NETWORKS

YALÇIN, ÖZGE NİLAY
M.S., Department of Cognitive Science
Supervisor : Prof. Dr. Cem Bozşahin

September 2014, 47 pages

In this study we proposed a model that highlights the effect of culture in language and form a hypothesis that suggests we can see these effects on the utterances of individuals and predict their behavior. We used a variation of the famous naming game to simulate our model, and later compared our results with the empirical data we collected from an online social network platform, Twitter. The use of hashtags as an act of labelling for popular topics is investigated due to the resemblance of the phenomenon with the naming game. The simulation of the model created a population with varying preferences on the topics of communication, which is a more realistic approach than the always converging case of the classical naming game. Another cultural force on partner selection, generated a topology within the population from an initial state of homogeneity. Empirical results were compatible with the simulation that uses those cultural forces and the model has a high predictive power within the scope of selected topics. A strong correlation is found between the similarity of previous use of hashtags and the hashtag use about two important political events. However, the model requires a large number of data for implementation and also has computational limitations that makes it difficult to be used for practical purposes.

Keywords: semiotic dynamics, naming game, communication theory, social networks, evolutionary linguistics

ÖZ

DİL OYUNLARI VE SOSYAL AĞLAR KULLANILARAK KÜLTÜRÜN
İLETİŞİM ÜZERİNDEKİ ETKİSİNİ MODELLEME

YALÇIN, ÖZGE NİLAY
Yüksek Lisans, Bilişsel Bilimler Bölümü
Tez Yöneticisi : Prof. Dr. Cem Bozşahin

Eylül 2014 , 47 sayfa

Bu çalışmada öne sürülen model kültürel kuvvetlerin dil üzerindeki etkilerini göz önüne alarak, kişilerin dilsel davranışları üzerine tahmin yürütmeyi amaçlamaktadır. Modelin simülasyonu kapsamlıca incelenmiş bir dil oyunu olan 'naming game' modelinin bir varyasyonu kullanılarak yapılmış, daha sonra elde edilen sonuçlar online sosyal paylaşım sitesi olan Twitter'dan elde edilen verilerle karşılaştırılmıştır. Bahsedilen isimlendirme oyununa olan benzerliği nedeniyle Twitter'daki popüler konular hakkında etiket kullanımları incelenmiştir. Modelin simülasyonu başlangıçta homojen olan bir popülasyondan, topolojik bir yapıya sahip ve iletişim partneri ile iletişim konusu üzerine farklı tercihleri olan bir popülasyon oluşturmuştur. Elde edilen bu sonuç, dil oyunlarının yakınsayan toplumlarına ziyade daha gerçeğe yakındır. Empirik verilerden elde edilen sonuçlar simülasyon sonuçlarına uyum sağlamış ve seçilen konular üzerine başarılı bir şekilde tahmin yürütülebilmiştir. Kullanıcıların geçmişte kullanmış oldukları etiket verilerinin benzerliği ile seçilen iki popüler politik konu üzerinde kullanılan hashtagler arasında anlamlı bir korelasyon gözlemlenmiştir. Fakat, modelin gerektirdiği veri miktarı ve hesaplamalar ile analizlerde harcanılan zaman gibi kısıtlar, modelin pratik amaçlar için kullanılmasını güçleştirmektedir.

Anahtar Kelimeler: işaretsel dinamikler, dil oyunları, iletişim teorisi, sosyal ağlar, evrimsel dilbilim

aileme

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor Prof. Dr. Cem Bozşahin for his guidance, invaluable support and encouragement throughout my graduate study. I am also grateful to Cengiz Acartürk, Deniz Zeyrek Bozşahin, Murat Perit Çakır and Ayşenur Birtürk for their constructive feedbacks.

I am especially thankful to my dearest friends Burcu Verim and Burçin İçdem for their joyous company and patience to my endless nonsense, which helped me to easily get through the most miserable moments. I thank all my friends, Murat, Furkan, İrem, Erdem, all "Bizimkiler" crew and Kevin for they could cheer me up everytime I needed. I also would like to thank to all those work on the Informatics Institute for their warm morning greetings and friendly conversations with a cup of tea. And last but not least, I am especially indebted to my parents, for the love, trust and support they provided me throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

Communication is a crucial aspect of biological, social, economic and artificial organizations. As social creatures, humans tend to communicate with each other via different media, such as: vocal, gestural and written signals. Although we share many forms of communication with other species, the capacity to communicate symbolically is thought to be the characteristic behavior of humans. Despite the fact that the process of communication involves much more than linguistic signals, the field of communication studies often centers around the use of language. Language is a highly complex, dynamical means of communication that includes interactions on different time scales. Therefore, it is important to understand the origins of language and the processes that led to its emergence. A large variety of disciplines contributes to the study of language evolution: ranging from archeology to anthropology, linguistics, biology, neuroscience, psychology and cognitive science. The topic of the emergence of language is highly controversial, yet researchers seem to agree that it is emerged out the of interaction of three forces: socio-economic, cultural and biological (Kirby, 2002; Steels, 2011). According to this view, there is a circular interaction between those three forces that work on different time scales. The biological mechanism of individuals is shaped by evolutionary forces that establish the necessary structures to learn and produce a language. Social forces provide a setting for those individuals to interact in an environment that requires information transmission. Within this framework, cultural forces help to create a shared system within a population that allows an efficient communication.

Culture and language are highly interrelated properties of a society that shape and being shaped by the communicative actions between individuals. The construcitivist view of culture as a meaning making process further puts this relation in a more prominent place in the emergence of language within a community. Evolutionary linguistics is a research camp that studies the cultural forces that shapes the language.

One of the main challenges for the cultural examination of language is the difficulty of gathering empirical data. Most of the research on this topic is dedicated to extensive studies on computer simulations and robotics, with the aid of the techniques from semiotic dynamics.

Lately, computational simulations have become a widely used approach to study the emergent properties of language on these different time scales. Evolutionary approach that uses genetic algorithms focuses mostly on biological forces, iterated learning model focuses mostly on the socio-economic forces and evolutionary linguistics approach focuses on cultural forces. All of these perspectives have been extensively studied in recent years both in situated and non-situated environments.

Evolutionary linguistics tries to give an answer to the questions about the emergence of language by forming the hypothesis that language is primarily shaped by cultural forces (Steels, 2011; Croft, 2008). Agent based simulations form an appropriate basis to test this hypothesis. The interactions between agents lead to changes in information transmission over time, which in turn forms culture. The constructivist view of culture as a meaning creation process suggests that culture shapes and being shaped by the interactions between agents (Vygotsky, 1997). Over the past years many multi-agent simulations were created to examine the effect of culture on language emergence (see Steels, 2011).

One line of such models uses the idea of a language game, a concept that have been first proposed by Wittgenstein (1953), in order to develop intuitions about the underlying forces on language. Naming games are one prominent way of examining the cultural conventions that shapes language, that have been recently increasingly popular. However, the direct effect of culture into individuals' choice of conversation topic, have not been a part of the models.

In this study, we constructed a communication model that will take the effect of cultural forces on language into account and simulate it using an agent based approach of the naming game paradigm. We aim to see the direct effect of culture in the utterances of both artificial agents and humans in a real social network. Our hypothesis was that the more similar the utterances of the individuals up to a certain point in time, the more they will tend to talk about similar topics in the future. In order to test this hypothesis we used two methods, simulating an agent based model and analyzing real world data from the online social network: Twitter. In the following chapters a literature review will be provided on the concepts of communication dynamics, cultural dynamics and language dynamics which are in close connection with each other. Subsequently, detailed explanation of the methodology used in the study will be given both on the simulation and the empirical study, followed by the results and conclusions drawn from them.

CHAPTER 2

BACKGROUND

In this study, we will examine the effect of culture in human communication and form a hypothesis that predicts the hashtagging behavior on the Twitter platform. In order to reach our goal, we will first provide a literature survey on the concepts of culture and communication, followed by the studies on network dynamics and semiotic dynamics that are used in order to simulate the phenomenon. All of these will be held under the concept of language evolution, for the aim of this work is to provide an understanding on how linguistic conventions emerge, diffuse and change over time in populations and the two-way interaction of this process with culture.

Semiotic dynamics uses population of agents that interact with each other in order to reach a consensus on a language system. Within this framework, we can talk about two important aspects that shapes language and being reshaped in return. One of them is the act of communicating, where the other is culture as a meaning making process. In this chapter we will explain the concepts of communication, culture and the close connection between them. In addition, we will give two well studied methods (semiotic dynamics and social network studies) that are used in order to explain this phenomenon and give examples on some of the important work on these subjects.

2.1 Communication, Culture and Language

As being the main component of human interaction, communication occupies a significant place in social science for many years. Social interactions build language by effecting individuals as a result of aligning mental representations (Loreto et al., 2011). Language, as a symbolic form of communication is a unique behavior of humans that has been a subject of interest in cognitive science. Recent studies on the emergence of language and meaning making process formed the idea of a community of language users as a dynamic organization whose components develop a shared communication framework by interacting with each other (Loreto et al., 2011). In this chapter we will provide a literature survey on the concepts of communication, culture, language dynamics and their continuous interaction.

2.1.1 Communication Dynamics

Although communication may occur between technological and biological domains other than humans, throughout this study we will use the term communication to refer to the human interaction. The act of communicating is under investigation of many traditional approaches such as rhetorics, semiotics, cybernetics, sociopsychological and sociocultural studies (Craig, 1999). All of those approaches formed a different definition to communication, where each of them posed distinct questions that addressed various problems within the field. However, the interdisciplinary nature of the field has been an obstacle on the way of reaching a consensus. A collection of different definitions from various disciplines can be found in the paper of Craig (1999).

A simple linear model of this idea is the well-known model by Shannon (1948). The goal of constructing this model was to create a theory in order to formulate an effective way of transmitting signals from one location to another. The model puts information at the center of the communication process where the aim is for the receiver to reproduce the message the sender wants to communicate via a medium. As it can be seen in Figure 2.1, the information source that forms the message that is processed in the transmitter into a signal which will flow through a channel. During this transmission the signal might be exposed to a noise and finally reach to the receiver where it is decoded to an approximate message that will reach to the destination.

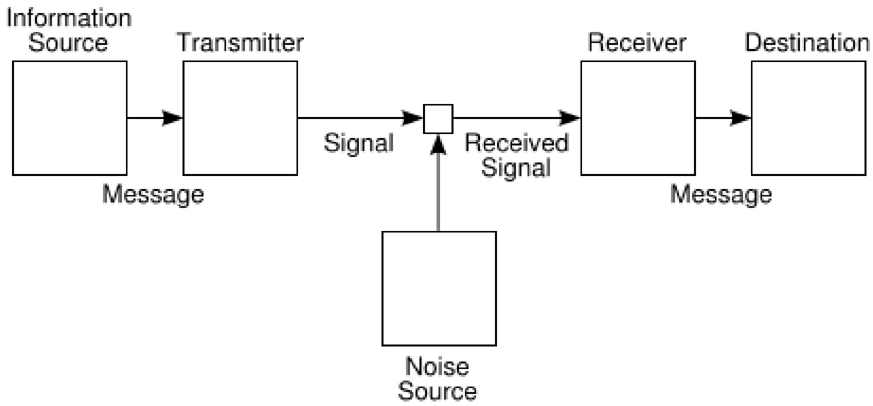


Figure 2.1: Communication Diagram by C. A. Shannon. From "A Mathematical Theory of Communication" by Shannon, C. A. 1948, *The Bell System Technical Journal*, 27, p. 381.

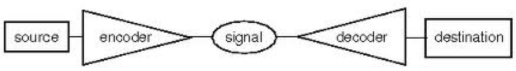
This mathematical model further developed into information theory, and gained vast recognition by solving various engineering problems in communication. Nonetheless, while trying to apply information theoretical approach to human communication, this model lacks a few components. First of all, there is no internal processing of the receiver and sender, therefore there is no connection of the messages to the meaning they refer, the model lacks the semantic connection. Consequently there are no individual differences that the model accounts for such as cultural traits, social skills and values, knowledge and so on. Additionally, as a linear model, it does not account for the feedback and dynamic properties of a conversation where cognitive science is more interested in.

Regarding these issues, Shannon and Weaver (1949) formulated three fundamental problems of communication:

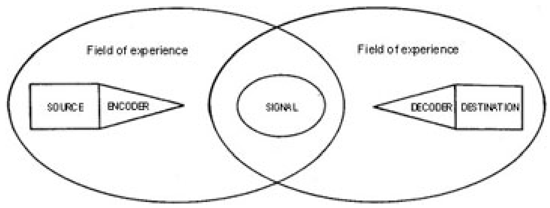
- the technical problem of how accurately the symbol is transmitted
- the semantic problem of how precisely the symbols evoke the desired meaning
- the effectiveness problem of how effectively the evoked meaning affect the referee in the desired way.

This divides the communication process to three levels. Their model solves the technical level with great success while leaving the other two unattended. This informational model of communication model has been and still being used by many scholars, as well as heavily criticized by many others (Craig, 1999). However, information theoretic approach is still the most widely accepted, most general and the simplest model available so far.

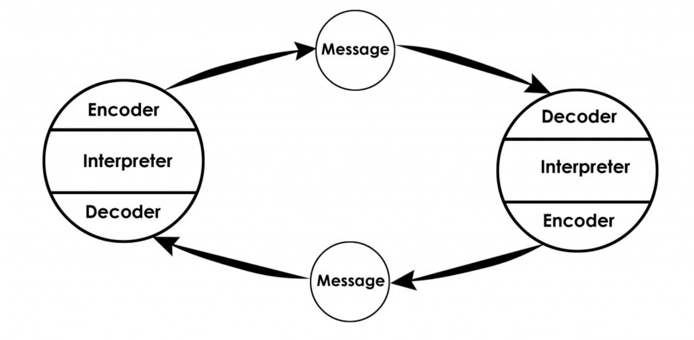
Lasswell (1948) defined communication as the process of "Who says What to Whom in What Channel with What Effect", which divides the process in five distinct components. This model highlights the effect the process invokes to the communication partners, regarding the third problem of communication. Communication, is an interaction having many effects on small and large scales. As flow of information, it affects mental processes of an individual by means of biological capacities of the human brain, such as learning, relating information with each other and reasoning. It shapes belief systems, norms and language of the individual, which in return shapes how to communicate further. This circular interaction between an individual and its environment is provided by communication. Consequently, as it progresses to a dynamic form of conversation, human communication involves more processes than just transmitting information.



(a) Simple Linear Model



(b) Field of Experience



(c) Dynamic model with Feedback

Figure 2.2: The communication diagrams by Schramm. From "How communication works" (p.3-26) by Schramm, W., 1954, In Schramm, W. (Ed.) , *The process and effects of mass communication*, 1954, University of Illinois Press, Urbana, IL.

An adaptation of Shannon and Weaver's model is Schramm's communication model, where he talked about the semantic aspects of the message (Schramm, 1954). He highlighted that there is no meaning in a message, unless people agreed upon some shared context. Starting from a simple linear model of communication similar to Shannon's, Schramm added a shared "field of experience" to the model where the communication partners' beliefs, norms, experiences and learned meanings provides a basis for the act of communicating (see Figure 2.2). Schramm suggested that without this shared space, two people cannot communicate. Moreover, he provided a circular interaction, where the communication is continuous (Schramm, 1954). The aforementioned components of the shared "field of experience" is highly similar to the elements that compose a culture.

Berlo (1960), like Schramm, is another scholar who highlights the effect of communication over culture in his communication model. In his Source-Message-Channel-Receiver (SMCR) model (see, Figure 2.3), he focuses on the past experiences that the communication partners as well as communication skills and social skills. This model individualizes the communication process and focuses on the differences the communication partners possess, suggesting a successful communication may not take place in a setting that partners shares none of those features.

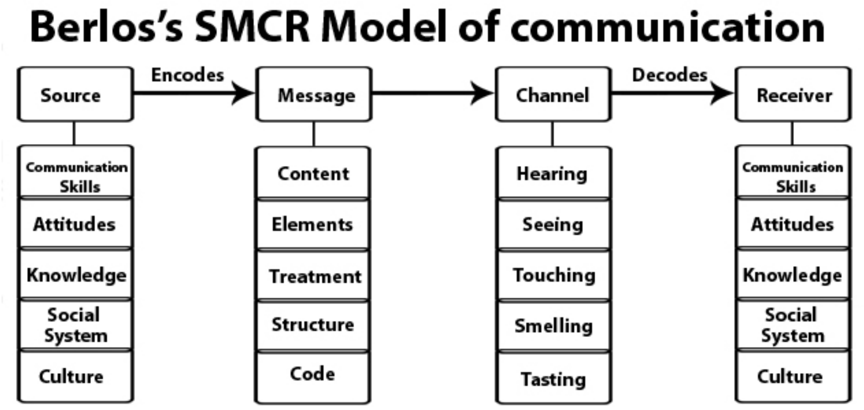


Figure 2.3: Communication Diagram by Berlo. From *The process of communication: An introduction to theory and practice*, by Berlo, D. K., 1960, Holt, Rinehart and Winston New York.

Blending these ideas on communication we may deduce that as being the member of the same species, we perceive the world in sufficiently similar ways that allows us to construct a link between shared signal and meaning pairs. All of the semantic connection is linked with communication by the shared experience, and this process shapes further communication patterns by constructing knowledge systems. These knowledge systems are constructed by denotational semantics, and it shapes what a message invokes in a person as well as how an individual may act upon a certain meaning. Both culture and language have this tight connection between the knowledge systems that is build by various forms of communication. This may be how individuals cope with the uncertainty of a given message (entropy in Shannon's terms) by constructing shared knowledge bases that are agreed upon. A semantic perspective on culture is therefore densely linked to language as both being the outcome of the communicative interactions within the context of shared experiences.

2.1.2 Cultural Dynamics

Culture, as a meaning creation process have gained much attention in cognitive anthropology (Bender et al., 2010). This line of thought suggests a two way interaction between culture and individuals, which is sustained by communication. This approach stresses out the importance of social influence within a given society and how culture shapes and being shaped by communicating individuals. It further proposes that communication becomes meaningful to us only if the link between events and experiences to certain values are generated by the cultural processes. Following this concept, cultures are thought to be constructed systems of meaning, that rise from the interaction between individuals. The effect of culture as a web of meaning constructed via social interactions has been gained favor among other scientists (Lull, 2002).

Cultural dynamics studies the cause, tenancy and change of culture, which is driven by dynamics within the population. These dynamics are a product of distinct behavioral forces that influences how a communication takes place, who will communicate and what effect will it have on the communicating individuals as we have mentioned in the previous chapter. As a group of individuals interact, the meaning space shifts, causing change in culture. Although it seems like the interactions will always brings order to a population of differing individuals, social forces may not always lead to a convergence in a population.

These forces are modeled in various studies to explain diverse social phenomena such using different tools such as opinion dynamics (Hegselmann and Krause, 2002), crowd dynamics (Bellomo and Dogbe, 2008), and language dynamics (Steels, 2011; Cangelosi and Parisi, 2002) (for a detailed review on agent-based models on social dynamics see, Castellano et al. 2009). A highly influential model of cultural dynamics is Axelrod's agent-based cultural model, where he defines culture as a set of attributes that may be influenced by social interaction (Axelrod, 1997). Within the scope of this study we will also refer to culture in Axelrod's sense and further investigate the cultural forces that drive diversity within a population.

In their recent paper, Kempe et al. (2013) investigate two basic forces; selection and influence, whose interaction drives cultural diversity. The force of influence is the tendency of people to effect the ones they interact with, which causes homogeneity in the community. The force of selection, on the other hand, can be defined as the tendency to interact with the people which are already similar, which causes fragmentation within the community. This selection force can also be named as homophily. There have been reported in many psychological literature that people tend to interact with similar others (Huston and Levinger, 1978). This is an important characteristic of social networks that is named homophily. A review of this phenomenon can be found in the paper of McPherson et al. (2001). Homophily is studied on real social networks (Singla and Richardson, 2008), revealing people that are in more interaction tend to be more alike and the more people interact the stronger the realtionship is. Social influence is also a recognized driving force in social networks. The power of this force depends on various factors such as strength of the relationship, population characteristics, individual differences, temporal and spatial effect (Sun and Tang, 2011). The effect of influence have been examined in large-scaled social networks in the study of Singla and Richardson (2008). Acting together, these two forces are claimed to be forming the cultural dynamics in a population.

Language is directly affected by this collective process of meaning making, as it provides a mapping between meanings and symbols. This close connection between language and culture have been the subject of the studies on emergence of language.

2.1.3 Language Dynamics

Language is a cognitive ability that is unique for humans that has a dynamical structure that includes interactions on different time scales. The emergence of such a complex system is under investigation by many scholars. The topic of evolution of language is highly controversial, yet researchers seems to agree on one aspect of it: that is it is emerged out the of interaction of three complex systems: learning, cultural evolution and biological evolution (Kirby, 2002). This topic clearly have an interdisciplinary nature ranging from anthropology to linguistics, genetics to robotics and each discipline have different theoretical and methodological views. Evolutionary linguistics tries to give an answer to the questions about the emergence of language by forming the hypothesis that language is primarily shaped by cultural forces (Steels, 2011).

Christiansen and Kirby (2003) states that language is the result of three distinct, yet interacting complex adaptive systems: individual learning (ontogeny), cultural transmission (glossogeny) and biological evolution (phylogeny). According to this view, there is a circular interaction between those three time scales, where the biological mechanism of individuals is shaped by biological evolutionary forces that affects their learning and production mechanisms. Learning, in turn shapes the cultural transmission of language in a population of language users. This complex interaction model can be seen in Figure 2.4.

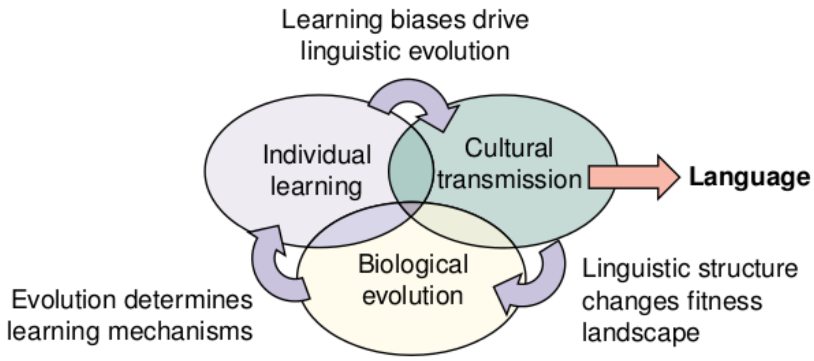


Figure 2.4: Interactions of language dynamics on different time scales. From "Language evolution: consensus vs controversies" by Christiansen, M.H. & Kirby, S. 2003, *Trends in Cognitive Sciences*, 7, p.302.

Another view on language evolution that also defines language as a complex adaptive system is pioneered by (Steels, 2012), focuses on the primary forces that shapes the language as: biological, social and cultural forces. In this view, the biological structures are the innate mechanisms that shapes language universals which evolved in human primates some million years ago, must be explained via evolutionary forces. These evolutionary forces will result in increased brain capacity that forms the basis for a language to emerge. Additionally, language is a social phenomenon that have emerged from collective activity. The case of Nicaraguan sign language shows us, a shared language can not emerge without a shared context and social interaction (Senghas et al., 2004). In a society where societal and ecological complexity increases, there will be social forces that affects the language. Biological evolution supports the social evolution in this context. Also, increased complexity in society requires increased linguistic complexity in order to satisfy the need to interact with that highly complex environment. This complex interaction model is summarized in Figure 2.5. Continuous lines in the figure shows how increased biological complexity will effect how individuals interacts with the world, therefore result in an increased social and ecological complexity which also supports linguistic complexity. The dashed lines show "requires" relation, as an increase in linguistic or socio-economic processes requires an increase in biological complexity.

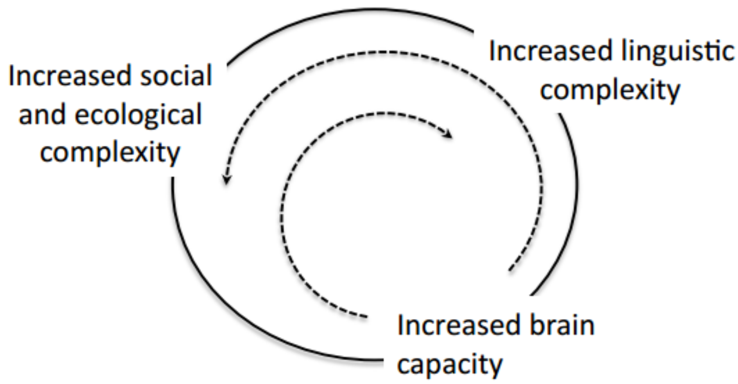


Figure 2.5: Interactions of forces that shapes language.From "Self-organization and selection in cultural language evolution" by Steels, L., 2012, in *Experiments in Cultural Language Evolution*, John Benjamins, Amsterdam.

The interaction between cultural and linguistic forces as they both are shaped by communications between individuals, have been under investigation by using multi agent simulations. Communicating individuals as being modeled by agents in a computer environment provides a basis for the understanding of how much linguistic structure we can capture with sociocultural forces. Moreover, social network studies adds further insight to how real life relations shapes and being shaped by those interactions. In the next section, we will investigate these methods in depth.

2.2 Methods on Communication

2.2.1 Agent-Based Simulations

It is important to understand how the interaction between individual components give rise to global effects that was not present in lower levels. Agent based simulations have been used to show this emergent property in many scientific domains ranging from biology, physics to social sciences and economics (Macal and North, 2008). An agent based system is composed of simple interacting components, which is called actors or agents, that have certain features that allows them to act in their environment. This environment may consist of static elements such as objects, as well as elements that can act upon and change their structure, such as other agents. Social features are modeled in terms of a set of variables that is subject to social interaction. These features can be defined as binary variables which is widely used in opinion dynamics, set of object-word pairs where we can see in the example of language dynamics or matrices in the case of cultural dynamics and belief systems. In all of these approaches the social dynamics tend to move towards a shared feature space (whether it is a lexicon, culture or opinion), as a direct result of those forces.

This approach provide valuable insight to the understanding of complex population behavior that emerges from those simple interactions of actors. In social sciences, where the actors are humans, agent based modeling provides a basis to the study of social dynamics, where the aim is to explain the macro-level structures from the micro-level interactions (Castellano et al., 2009). This bottom-up approach is widely used to explain various kinds of social phenomenon such as crowd behavior models (Helbing, 2001; Epstein, 2002), opinion dynamics (Hegselmann and Krause, 2002; Galam et al., 1982), cultural dynamics (Axelrod, 1997; Klemm et al., 2003) and language games (Steels, 2011; Kirby, 2002; Baronchelli et al., 2008; Loreto et al., 2011).

Recently, language dynamics has been studied in computer simulations, mostly in agent-based models, in forms of naming games. Agent-based simulations forms a basis to investigate the forces that shapes the language. Computer simulations provide a more controlled system, where we can see the effects of the assumptions in a model. One line of research on the emergence of language focuses on the social use of the language in order to explain its characteristics. A shared lexicon is one of those characteristics that is claimed to be shaped out of social interaction. This line of research is pioneered by computational simulations, namely multi-agent simulations where a population of agents interact in a simulated environment. Computational simulations use the language game concept as a basis for language evolution research. Although culture is an important part of meaning making process, these models did not dwell into the cultural aspects of language. The ability to influence other agents is one of the important cultural forces in agent based simulations, where selection can be named as the other (Kempe et al., 2013). In this study, i will focus on the effect of culture on linguistic behavior by using these forces on a language game setting.

2.2.2 Language Games

The concept of language games are first proposed by Wittgenstein (1953), where a language serves for providing a communication between two agents by referring to the objects in their environment with signals. He proposed to use this model as a primitive language, to have a better understanding on linguistic interaction. It is important to see that language games have little similarity with real world language, however as Wittgenstein puts it it is important to understand the similarities and dissimilarities between those models and natural languages in order to have a complete understanding of language structures (Wittgenstein, 1953).

Language games differ according to their communicative goal, population of individuals and a context (Steels, 2012). The naming game, which is a specific case of language games, resembles the language game concept of Wittgenstein in terms of communicative goal and context. Computational simulations of the naming game, models a population of agents that interacts with each other using a specific case of language games. Agent-based simulations of the language games can be situated and non-situated (Wagner et al., 2003). Situated simulations requires the agents to form their own perceptions about the environment, where non-situated agents are given pre-defined meanings that is defined by the researcher. Although the situated approach is more realistic in terms of language production and acquisition (Hutchins and Johnson, 2009), non-situated simulations provides a simple environment that a language can emerge or be learned. In this study, I will model the naming game with non-situated simulations in order to examine the social forces that shapes a language.

2.2.3 Naming Game

The naming game is a specific case of language games where a population of agents tries to form a shared lexicon (Vylder and Tuyls, 2006). In a naming game, agents assumed to have prior biases. Although the nature of those biases may vary between studies, naming games in a non-situated simulation environment have similar assumptions (Wagner et al., 2003). In the naming game it is assumed that the agents knows how to send, receive and discriminate between signals. Note that, this assumption tries to cover the minimal cognitive mechanisms that are related to language production and comprehension. It is also assumed that the objects in the simulation environment are identifiable by each agent. In addition to those assumptions, it is further assumed that the agents have a motivation to communicate with each other using those cognitive abilities and the aid of an independent communication channel, such as gestures, in order to confirm or point out the meanings they refer to.

The original naming game consists of a population of agents, all of which can send and receive a signal. Figure 2.6 shows the naming game, where a sender and a receiver is selected from a population of agents and communicate about the meanings in the environment. The communication will be evaluated in order to its success.

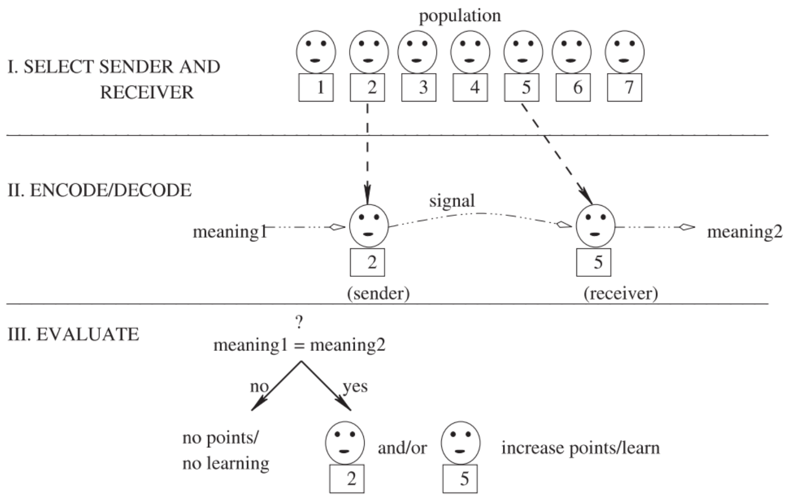


Figure 2.6: Naming Game Diagram. From "Progress in the simulation of emergent communication and language", by Wagner et. al., 2003, *Adaptive Behavior*, 11, p.37-69.

The outline of the game can be shown:

1. Two agents are chosen randomly from the population. One is assigned as speaker, the other as hearer.

2. Speaker chooses an object to communicate about (the topic), and selects a proper name for it from its lexicon. If the corresponding name is not present in the lexicon of the agent, the agent produces a novel signal in order to refer to that object. Speaker utters the word it selects for the object to the hearer as an act of communication.
3. Hearer matches the communicated word to a meaning in its own lexicon, and gives feedback to the speaker using another communicating channel (e.g. pointing).
4. According to the success of the communication, both agents update their lexicons. A language strategy will determine how the agents update their lexicons.

In a naming game, which language strategy the agents will use after a successful or unsuccessful communication is important. Various language strategies have been studied throughly in numerous studies. One strategy of a naming game holds the word-object associations as a weighted list. In a generalized model of the naming game proposed by (Vylder and Tuyls, 2006), agents do not clear their lexicons entirely after unsuccessful communication but update their inventories according to a weighted distribution. Another strategy of lateral inhibition was introduced to the model by Steels (2000), mimicking the natural altering behavior of excited neurons over the action potential of neighboring neurons. The idea behind lateral inhibition strategy is when a successful communication occurs with a certain object-word pair, every other word related to that object will be inhibited. A model by Steels et al. (2005), introduces three parameters to the game that functions on the lexical entries: $\delta_{increase}$, $\delta_{decrease}$ and $\delta_{inhibition}$. These parameters alter the word scores on the lexicons of the agent; $\delta_{decrease}$ decreases the weight in an unsuccessful game, $\delta_{increase}$ increases the weight of the selected object-word pair in a successful game while inhibiting the other words associated to that object by $\delta_{inhibition}$. The minimal meaning game is the basic mean field case of the naming game models, where there exists no topological structure (Baronchelli et al., 2008). In the minimal naming game, if the receiver's signal-meaning pair will not match with the sender's signal-meaning pair, the receiver adds the new signal to its lexicon. If the communication is successful, both agents deletes their entries for the corresponding object for all but the successful signal. This model is actually a special case of the lateral inhibition strategy with three parameters.

Note that in all cases the topics have equal probability of getting selected, as well as the agents. Social network methods are highly beneficial tools in order to study the topological properties which strongly effects how social interactions effect naming game dynamics. In order to investigate these properties we used network analysis methods, which we provide a review in the next chapter.

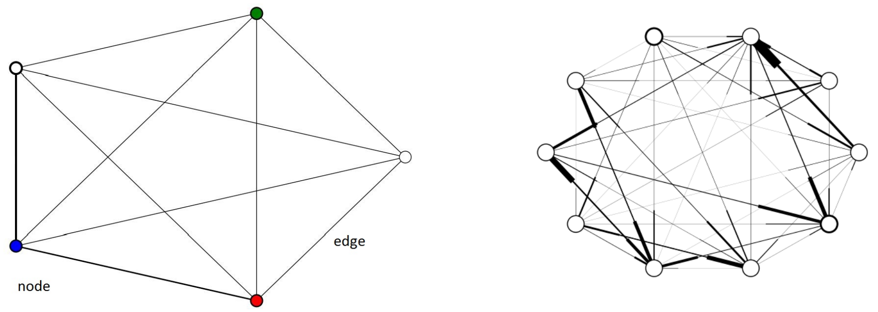
2.2.4 Social Networks

A social network is any network where the nodes are humans, and the edges are the relationship or interaction between these individuals (Aggarwal, 2011). The form of those interactions may involve an information transfer on various mediums, such as text, voice or gestures; while the relationship it constitutes might be kinship, friendship, follower-followee relation etc. The study of networks are derived from mathematical graph theory and has been using its terminology. In social sciences the study of networks is widely used to examine the concepts like community structure, centrality or popularity of nodes, and connectivity of a network.

The mathematical investigation of networks takes places under the study on graph theory. In graph theory, a network can be represented by a graph by agents as nodes, and the relationships between agents as edges. The notation of the graph theory is as follows:

1. A set of nodes, or vertices $\mathcal{N} = n_1, n_2 \ldots n_k$ where l is the number of nodes.
2. A set of edges, or lines $\mathcal{E} = e_1, e_2 \ldots e_l$ where e is the pair of nodes that the link exists in the form of $< n_i, n_j >$ and l is the number of edges. In a non-directed complete graph, the network will have $l = k - 1$ edges in total.

The overall graph is notated as $\mathcal{G} = (\mathcal{N}, \mathcal{E})$. In Figure 2.7a, a mean field network of five nodes can be seen. Nodes are represented by circles, where edges are the lines between them. In this particular network, the strength of the edges are represented by the thickness of the lines. Additionally, Figure 2.7b is a directed network with ten nodes, where each edge have at most two directions.



(a) Non-Directed Network

(b) A Directed Network

Figure 2.7: Networks with various properties.

The relationship may be one way (directed) or two-way (nondirected), as well as it could be static or dynamic. Static networks does not change due to the relationships or time, where dynamic networks continuously change. The edge properties may vary due to the relationship one want to capture or focus in the graph, and it may have weights, ranking or types, as well as other kind of symbolic representations. Graph theory, as a branch of mathematics, have been extensively studied and it is widely used in social network studies.

As a well studied topic, it has been valuable for social and complex network studies. A famous example of a social network is the study of Milgram (1967) on the well-known small-world phenomenon(also known as six degrees of separation), that hypothesized any person in the world is linked to the other with an average of six other acquaintances. This small world property have been found in many biological and social networks, as well as power-law degree distributions and network transitivity (for an extensive review, see Newman 2003). In terms of social networks the property of network transitivity, also called clustering, suggests that if two person in a network are connected to each other while one of them is connected to a third person, it is highly likely that the other node will be connected to the third person as well. Clustering coefficient shows the degree of the connectivity of the nodes. Social network analysis have been recently used to examine the interplay between cognition and social structure (Baronchelli et al., 2013), adoption of technological innovations, social influence (Kempe et al., 2003; McPherson et al., 2001), community formation (Girvan and Newman, 2002).

It is a well studied phenomenon both mathematically and sociologically that, the overall structure of the network connections has a high impact on spread of information (Watts, 2004). The structure of a network that shows the interaction characteristics between nodes is called the topology of that network. It may show how intense, frequent or probable an interaction is. The topology of the network structure highly influences the information distribution, therefore changing the convergence behavior of the system. The topology of the network might change in time due to interactions between nodes, or it might show a static property. Network topology impacts the spread of information within a network, that results in variations in the community structure and convergence dynamics of the population.

The mean field case is the base case of networks where each agent is connected to each other. In naming game studies generally use mean field topology, each agent have equal probability to get selected as speaker or hearer. However studies on social networks showed that natural network topologies usually show small-world properties similar to various types of complex networks such as biological and technological networks (Newman, 2003). However, there exist some differences between social networks and other types of networks in terms of assortativity and clustering coefficient properties (Newman and Park, 2003). Assortativity of a network is a property that refers to the connection preference of nodes favoring to interact with nodes that have similar degree with itself. Clustering coefficient of social networks seem to be higher than the average random network, suggesting that it is much more likely for two nodes to be connected if there exists a connection between their already connected partners than it is expected from other types of networks.

The role of topology on the agreement dynamics in a naming game setting studied in the paper of Dall'Asta et al. (2006). They found that the order of the selection of the conversation partners effects the number of different words, average success rate and convergence time (Baronchelli et al., 2006). In the direct naming game case where the speaker selects a conversation partner among its connected agents is considered to be more natural than the neutral case or the reverse case where hearer selects a conversation partner (Barrat et al., 2007). Therefore, we will use the direct naming game model in our study.

The topological structure can be implemented in a naming game setting on a network model. It has been shown that networks with small-world characteristics show faster convergence and no community structure (Lu et al., 2009). They also examined the naming game in real social networks, in this case a friendship network and saw the population does not reach a consensus and there are several communities that coexist. These studies were conducted on networks with a static topology. Properties of a naming game with a dynamical network property are examined in the paper of Lipowska and Lipowski (2012), where they build a model of weighted network structure. In their model, the edge weights depends on how successful the previous communications between agents were and they have found a multi-lingually stable community structure in some of their parameter settings, more similar to real world phenomenon. In our study, we will model naming game on a network of agents that obey to the selective cultural forces which we previously discussed, that results in a similar dynamical topology with this study. In the following sections we will explain the detailed methodology of our study.

CHAPTER 3

METHODOLOGY

3.1 The Model

The proposed model in this study is a modification of the naming game that differs from the classical model in terms of selection forces. This is applied in the model by the selection of the communication partners, and the topic to be communicated. As we have changed more than one property of the original model, to be able to compare and see the effects of both processes we have formed the model gradually. First, we have replicated the original naming game with weighted lexicon properties. In this model, any agent have equal probabilities for getting selected. Additionally, every object have an equal probability to get chosen as a conversation topic. This will lead to a convergence in terms of lexicons, therefore a homogeneous community. In order to see the effect of culture, we will add a selection force that changes how an agent choose a communication partner. In this scenario, the agents will tend to interact more with the agents that they have interacted successfully before. This means that the agents will need to keep the scores of how successful previous conversations were with every other agent. After comparing the results with the original model we will add another selection force that effects how the conversation topic is determined. In this final model, the agents will tend to talk more about the topics they have talked before. We expect this selective preference will lead to a cultural divergence within the community. It should be kept in mind that no prior topology in the model, but a topological structure is being formed out of the selection preferences of the agents.

3.1.1 The Simulation Environment

The model is simulated on computers using various hardware and software settings. Simulations are implemented with *Python* programming language, using *Numpy* library for scientific computing, *Networkx* for creating the network graph, and *Graphviz* and *Matplotlib* libraries for visualizations and plotting.

The simulation environment consists of a group of agents, that can interact with each other about the objects that are situated in the environment. In the following sections, we will provide a detailed description of each of those elements.

1. **Objects** There are a number of objects that are situated in the environment $o_1, o_2 \ldots o_n \in \mathcal{O}$. Agents form their lexical properties and topic lists according to these objects.
2. **Agents** A population of agents $\alpha_1, \alpha_2 \ldots \alpha_m \in \mathcal{P}$ are situated in the simulation

environment, where *m* is the size of the population. Each of the agents have a distinct lexicon, a signaling system that allows them to communicate and internal functions that manipulates the lexicons and signals.

(a) **Lexicons:** Each agent have a lexicon $\mathcal{L}_{\alpha_i}$. The items in the lexicon consist of objects and names associated to that object, where each of them will have a weight. The lexicon of an agent can be altered according to the communicative actions the agent gets involved.
A speaker chooses a word that is attached to the chosen topic in its lexicon, with the function $f_{chooseWord}$. It selects a proper word in the lexical item according to its probability. This probability is computed via roulette rule, where the weight of the word w_j for the object o_i is ω_{o_i,w_j} and the probability is:

$$p_{o_i,w_j} = \frac{\omega_{o_i,w_j}}{\sum_{j=1}^{k} w_j} \tag{3.1}$$

for a total of *k* number of words attached to that word for that agent. If there exists no word attached to that object in the lexicon of the agent, then the agent invents a new word for that object using f_{invent}. This function simply produces a signal with the pre-determined length, composed of lower case alphabetical letters.
Speaker later utters the chosen word to the selected hearer with f_{speak}. Hearer, interprets this word using $f_{chooseObject}$ by selecting the highest value the word has in its lexicon.

(b) **Topic List:** Each agent have a weighted topic list $\mathcal{T}_{\alpha_{o_i}}$, where each topic have a weight according to the communicative frequency of it. The topics are linked to the objects in the environment.
Each agent is capable of selecting a topic, according to the probability of it. This is computed using a $f_{chooseTopic}$ function, according roulette rule, where the probability of the topic o to get selected from the number of m topics (hence, objects) is:

$$p_{o_i} = \frac{\psi_{o_i}}{\sum_{j=1}^{m} \psi_{o_j}} \tag{3.2}$$

This will allow a topic to get selected even though it has a low weight.
If a topic o is chosen to communicate, the weight of that topic will increase by ε regardless of the success of the communication:

$$\psi_{o_i} = \psi_{o_i} + \gamma_{update} \tag{3.3}$$

(c) **Network Structure:** Initially the population topology have a mean field characteristic. Each agent holds a weighted list in its memory that shows the strength of the connections it has with its pairs:

$$\mathcal{W}_{\alpha_i} = \omega_{i,1}, \omega_{i,2} \ldots \omega_{i,j}, \forall \omega_{i,j} \text{ if } i \neg j \tag{3.4}$$

This weighted list gets updated according to the result of the communication between pairs.

(d) **Parameters:** In the model we have a highly dynamic structure in terms of lexicons, topic lists and network topology. There are minimum and maximum values for all weighted items, as well as an epsilon value that is used

for updating the weights according to the communicational success. The weight of a word w in a lexical item attached to the object o that is represented as $\omega_{o,w}$ will be updated with the parameter θ_{update}, which has the mininum and maximum values of $(\theta_{min}, \theta_{max})$. The weight of a topic for an agent ψ_o will be update with γ_{update} parameter regardless of the communicative success until the maximum value of (γ_{max}) is reached. The weight of a link between two agents α_i and α_j is represented as $\omega_{\alpha_i,\alpha_j}$, and it will be updated by δ_{update} parameter that has the minimum and maximum values of $(\delta_{min}, \delta_{max})$.
In a successful communication between the agents $(\alpha_i, \alpha j)$ of the object and word pair (o, w):

- both agents updates their lexicons:

$$f_{success}(o, w) = \begin{cases} \omega_{o,w} + \theta_{update} & \text{if } \omega_{o,w} + \theta_{update} < \theta_{max} \\ \theta_{max} & \text{else} \end{cases} \tag{3.5}$$

- both agents update the weight of the topic in their topic list:

$$f_{updateTopic}(o) = \begin{cases} \psi_o + \gamma_{update} & \text{if } \psi_o + \gamma_{update} < \gamma_{max} \\ \gamma_{max} & \text{else} \end{cases} \tag{3.6}$$

- the speaker updates the weight of the link between itself and hearer:

$$f_{successLink}(\omega_{\alpha i,\alpha j}) = \begin{cases} \omega_{\alpha_i,\alpha_j} + \delta_{update} & \text{if } \omega_{\alpha_i,\alpha_j} + \delta_{update} < \delta_{max} \\ \delta_{max} & \text{else} \end{cases} \tag{3.7}$$

In an unsuccessful communication between the agents (α_i, α_j) of the object and word pair (o, w):

- both agents updates their lexicons:

$$f_{failure}(o, w) = \begin{cases} \omega_{o,w} - \theta_{update} & \text{if } \omega_{o,w} - \theta_{update} < \theta_{min} \\ \theta_{min} & \text{else} \end{cases} \tag{3.8}$$

- both agents update the weight of the topic in their topic list as it is presented in (3.6). Note that the topics get updated the same way regardless of the success of the communication.
- the speaker updates the weight of the link between itself and hearer:

$$f_{failureLink}(\omega_{\alpha i,\alpha j}) = \begin{cases} \omega_{\alpha_i,\alpha_j} - \delta_{update} & \text{if } \omega_{\alpha_i,\alpha_j} - \delta_{update} < \delta_{min} \\ \delta_{min} & \text{else} \end{cases} \tag{3.9}$$

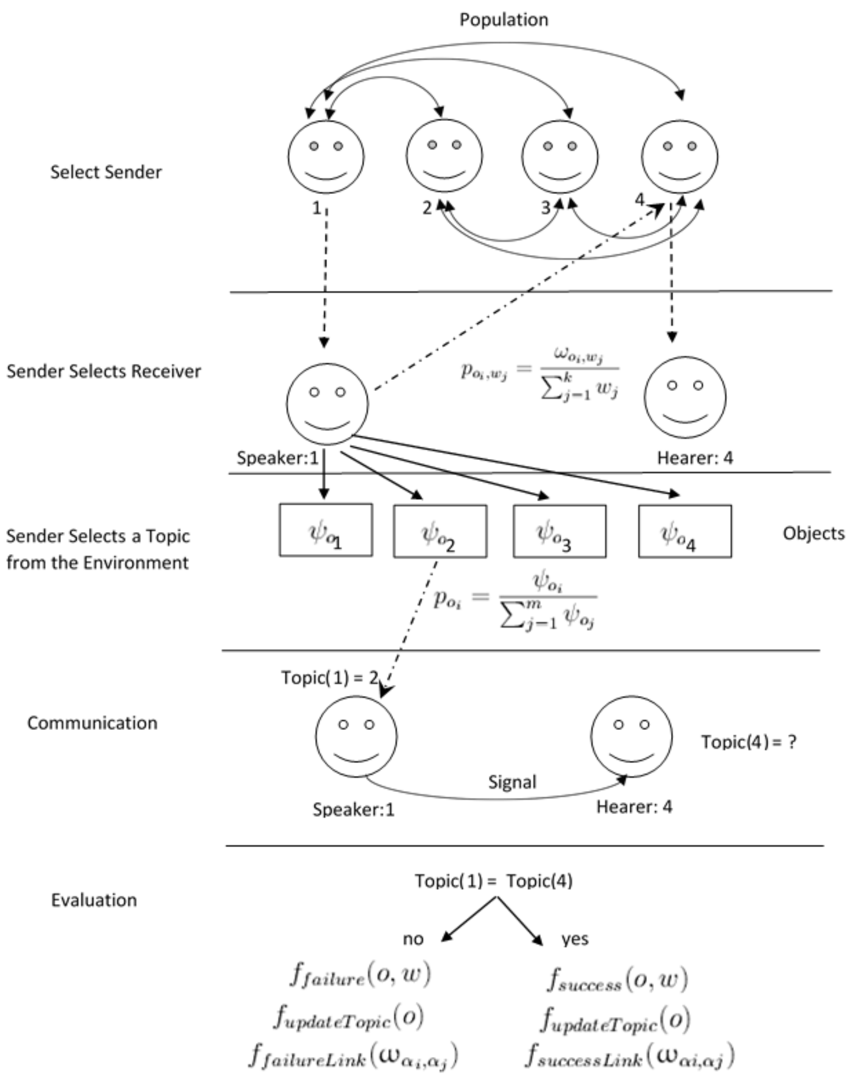


Figure 3.1: The diagram of the proposed model.

The proposed model can be seen in Figure 3.1, while a detailed diagram of the communication between partners can be seen in Figure 3.2.

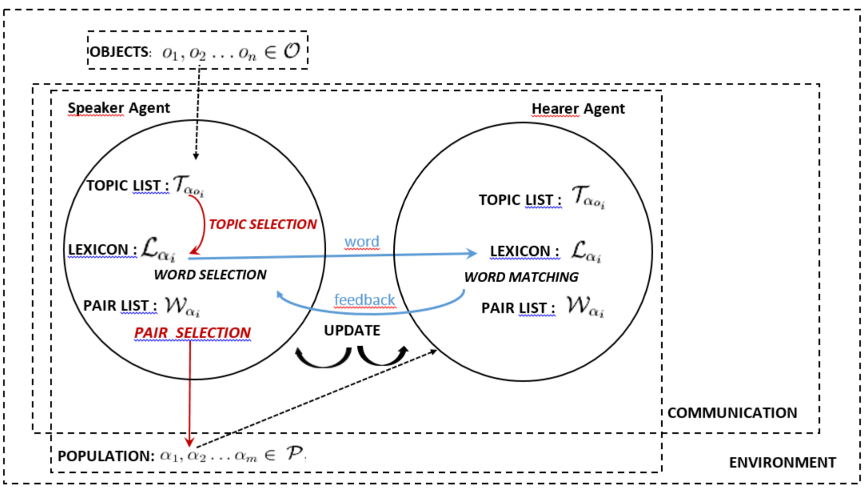


Figure 3.2: The diagram of the communication between pairs.

The simulation involves these steps in one iteration:

1. A speaker is chosen from a population of agents.
2. Speaker chooses the topic and another agent that it wants to communicate, according to their probability.
3. Speaker chooses a word for the chosen topic, or invents a word if such a word does not exist. Speaker sends the word to the hearer.
4. Hearer interprets the word by matching it to an object in its lexicon, if that word does not exist it chooses a random object and gives feedback to the speaker via a different communication medium.
5. In the case of a successful communication where the speaker and hearer both choose the same object:
   - both agents update their lexicons of the selected object-word pair using the function $f_{success}(o_i, w_j)$
   - both agents update the weight of the topic in their topic list using the function $f_{updateTopic}(o_i)$
   - the speaker updates the weight of the link between itself and hearer using the function $f_{successLink}(\omega_{\alpha_i,\alpha_j})$

   In the case of an unsuccessful communication, where the objects does not match:
   - both agents update their lexicons of their selected object-word pair using the function $f_{failure}(o_i, w_j)$
   - both agents update the weight of the topic in their topic list using the function $f_{updateTopic}(o_i)$
   - the speaker updates the weight of the link between itself and hearer using the function $f_{successLink}(\omega_{\alpha_i,\alpha_j})$

Note that the topic gets positively updated no matter how the communication results in. This is due to the assumption of the popularity of a topic is not associated with the agreement on the topic but its commonness among the whole population. In terms of language strategy, we used the weighted word list approach and investigate the outcome of such assumption where there is no lateral inhibition mechanism involved. Our model differs from the original naming game in terms of topic selection and pair selection mechanisms. We propose that these mechanisms will allow us to predict the agents with similar linguistic behavior, by directly effecting the lexical properties and cultural traits. The topological structure of the simulation model will be a directed graph, because the phenomenon we try to model is a directed social relationship between individuals.

First we used a classical naming game, with two parameters $\theta_{success}$ and $\theta_{failure}$, and analyze the interplay between those two parameters with a population of 10 agents and 5 objects, to be able to select an appropriate parameter set to further continue our study. We already know from previous studies that the convergence often does not occur in the case where $\theta_{success} \leq \theta_{failure}$. This effect can be seen in 3.3, where the convergence times for simulations where $\theta_{success} \leq \theta_{failure}$ took longer time to converge as shown by bigger and lighter circles. Therefore, we will discard those values from our later simulations.

Figure 3.3 shows time of convergence in each simulation of different parameters in a scatter plot, where $\theta_{max} = 1.0$ and $\theta_{min} = 0.1$, in a mean field case with no personal preference on topics . The x-axis shows $\theta_{success}$, y-axis $\theta_{failure}$ and the size of markers increases with the increasing convergence time where the darker the circle gets the lower the average time of convergence.

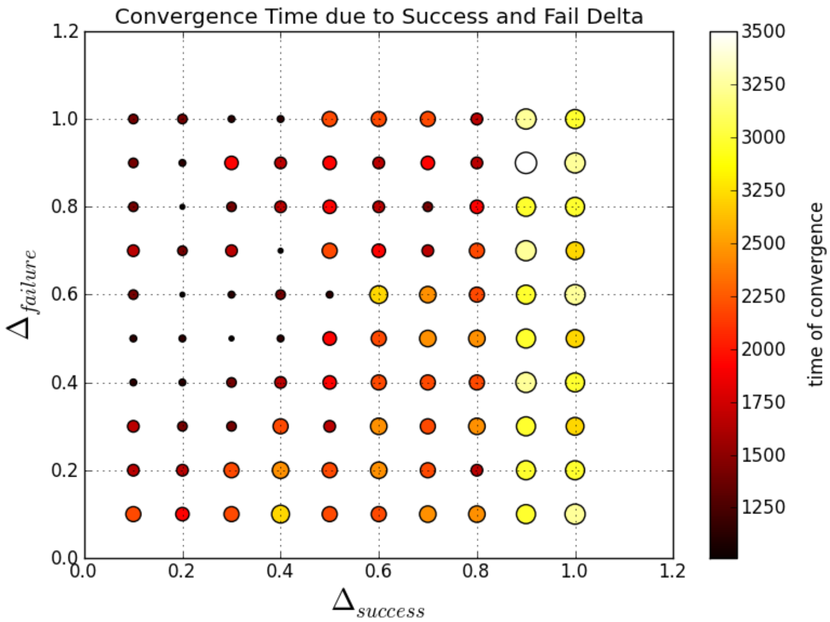


Figure 3.3: Time of convergence for different $\theta_{success}$ and $\theta_{failure}$ parameters.

We further examined the effect of the number of objects in the environment and the number of agents in the population on the success rates. In order to achieve this we used a parameter setting from previous findings, that allows the population to converge in a relatively short time. An increase in the number of agents further increase the time of convergence, therefore makes a population harder to reach consensus. However, the population still almost always have convergence on a shared lexicon. Figure 3.4 and Figure 3.5 shows the outcome of those simulations.

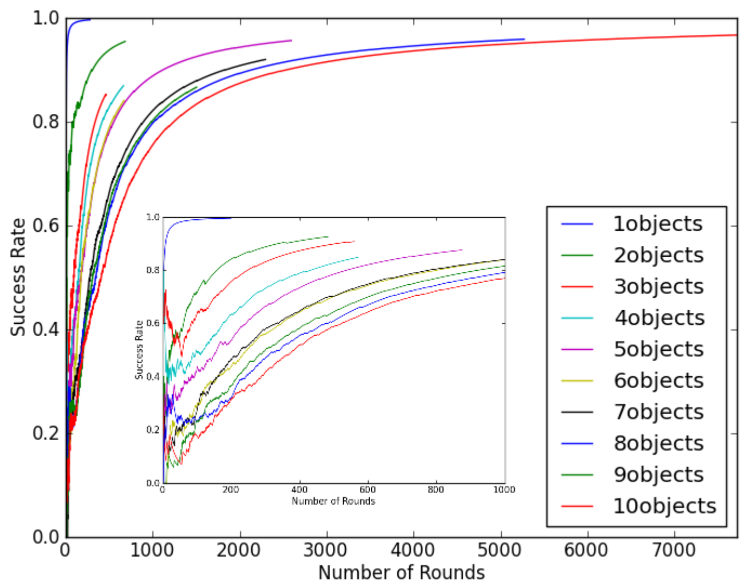


Figure 3.4: The change in time of convergence with an increase in the number of objects in the environment to be named.

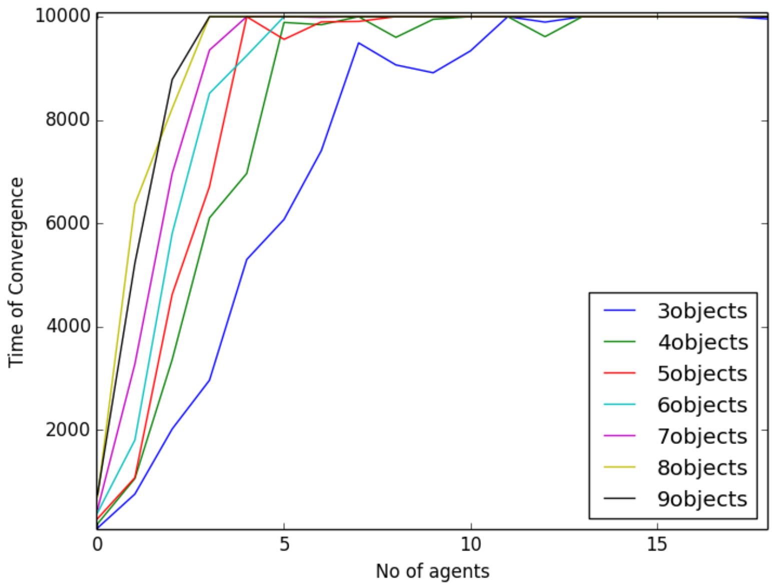
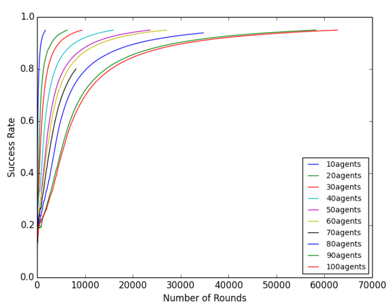


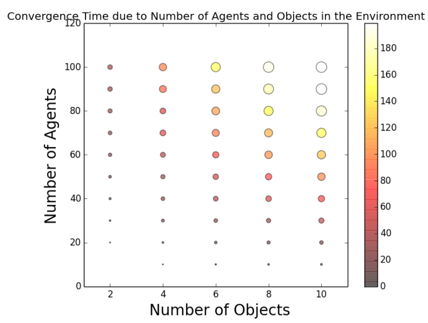
Figure 3.5: The change in time of convergence with an increase in the number of agents in the population.

As it can be seen from the Figure 3.4 and Figure 3.5, an increase in number of objects or agents results in an increase in convergence time. However, convergence occurs in all cases.

After examining the classical case, we used additional parameters to mimic the effect of culture as it has been described in the literature survey: the selection mechanisms. In order to model selection of communication partners as in the direct naming game case (Barrat et al., 2007), we represent our agents in a weighted network. Agents select each other according to the probability that is calculated considering those weights. In the study of Lipowska and Lipowski (2012), they showed that population converges to a single language regime with a fixed minimum value of the links between agents that allows the them to communicate even when they have unsuccessful communication. A simulation of this model shows similar and even faster convergent behavior with the mean field case.



(a) Success Rate due to the Change in Number of Agents



(b) Time of Convergence due to the Change in Number of Agents and Objects

Figure 3.6: Convergence dynamics of naming game on a weighted network.

As it can be seen in Figure 3.6, time of convergence shows the same behavior compared to the mean field case since all the agents have a chance to interact with each other.

If we change the minimum value for the weight of the links to be 0, allowing some links between the agents completely disappear after unsuccessful communications, the number of total languages used in the community severely changes. A sample community with 10 agents will look like Figure 3.7.

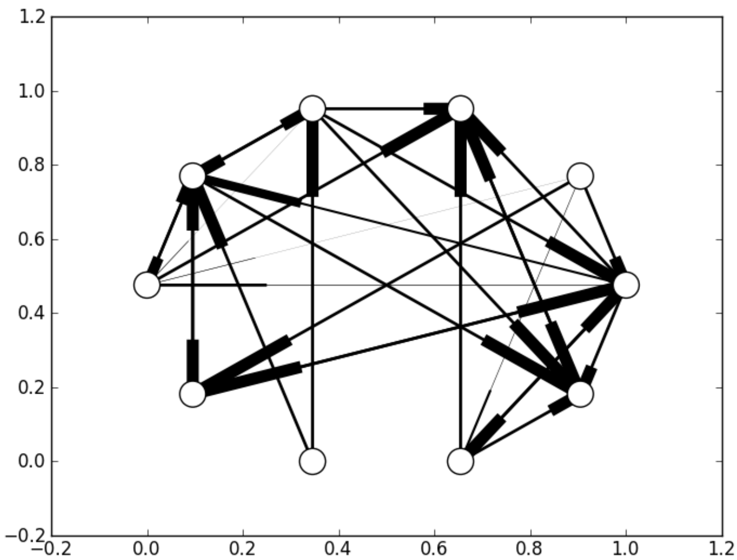
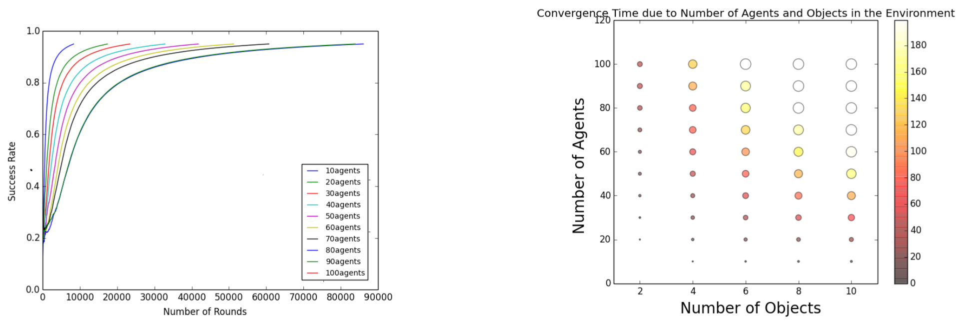


Figure 3.7: An example network of ten agents in the case of weighted links with a minimum value of zero.

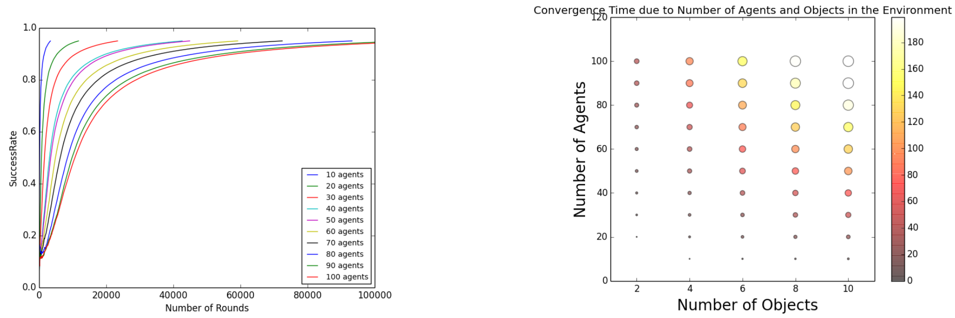
The convergence patterns and success rates are similar however in this case convergence takes a longer time in the network. The results can be seen in Figure 3.8.



(a) Success Rate due to the Change in Number of Agents **(b)** Time of Convergence due to the Change in Number of Agents and Objects

Figure 3.8: Convergence dynamics of a weighted network where the minimum value for the links are equal to zero.

The simulations so far were replicas of the previous studies, where the naming game was simulated in a social network with a selection mechanism that dynamically changes the topology of the network, and further result in a population with community structure. In addition to these to be able to simulate our hypothesis we added another selection mechanism that effects the internal processes of an agent: topic selection. Topic selection has been an area of interest for mostly opinion dynamics studies, rather than naming game studies. In our proposed model, we added a weighted topic list for every agent that represents their interest levels on each topic. Agents select the topic they want to talk about according to its probability, and the weights of topics get updated in each conversation. This list is used to show the effect of previous conversations, on the topic the agent will want to talk about in the future.



(a) Success Rate due to the Change in Number of Agents **(b)** Time of Convergence due to the Change in Number of Agents and Objects

Figure 3.9: Convergence dynamics of a weighted network where every agent has a topic list.

Finally we tested our hypothesis that the degree of two agents' similar past conversation topics will show the degree of them belonging the same cultural subgroup, therefore it will allow us to predict that they will tend to talk about similar topics in the future. To test this hypothesis we examined the conversation patterns of each agent in the last simulation setting, and tested for the prediction accuracy for two different scenarios.

Appendix-A shows all parameter settings used in order to test the hypothesis and the results according to the Pearson correlation r values and p values. The results showed a strong correlation between the lexicons and topic lists of the agents ($r_{lexicon-topic}$) for low p values ($p < .001$). Furthermore, Appendix-A shows the r and p values for the correlation between the similarity of words spoken for each object and the rest of the lexicon. For example for one object, we first calculate the similarity between the words used to describe that object for every agent pair and then look for a correlation between that similarity number and the words used for the rest of the objects for same agent pair. The correlation values were significantly correlated with a $p < .001$ 48 % of the objects, $p < .01$ 58 % of the objects and $p < .05$ for 70 % of the objects.

3.2 Twitter Study

In the second part of the study, in order to test our model in an empirical setting, we used an online communication phenomenon where individuals label the topics they want to speak about. These labels, which are called 'hashtags' are usually one word utterances identified with the '#' symbol that allows the users of social media to easily classify the content. In some online social platforms, hashtags are used to identify what is popular according to the number of individuals that uses that hashtag. In this study, we used the data extracted from Twitter platform, which is an online social microblogging website that allows its user to share and receive short texts or various electronic multimedia content within the limits of 140 characters. The reason for us to use this particular platform is the resemblance between the hashtagging behavior and the naming game. Additionally, we analyzed two special topics where the users show fast and competitive behavior which is perfect for an anology of a naming game.

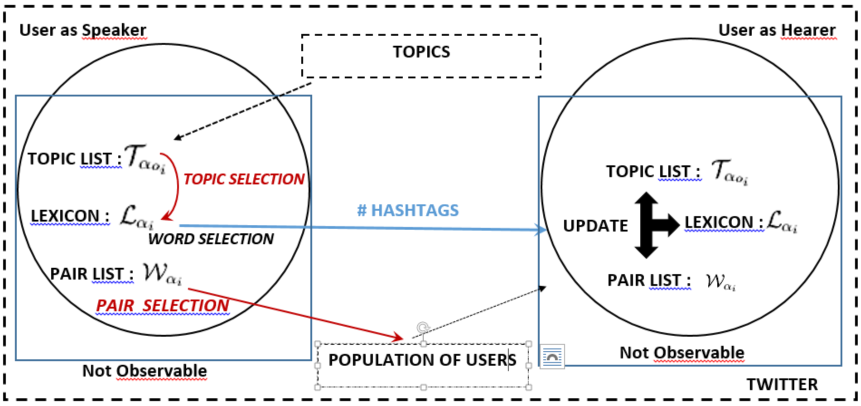


Figure 3.10: The analogy between the naming game and hashtagging behavior.

Figure 3.10 shows the analogy between hashtagging behavior in Twitter platform and the proposed model of the naming game. Unlike the simulation environment, we do not have direct access to the topic selection preferences and pair selection preferences of the users. However, the hashtagging behavior as a label of the preferred topic is similar to the naming behavior of the agents in the simulation environment where they label the objects in their environment with the names they created. Therefore we will examine the hashtag history of the users as an anology for the past utterances of the agents. This information will later be used to test whether there is a correlation between similarity of the past hashtag usage and the similarity of the hashtags on the selected topics, as it is proposed in the hypothesis. In this section, a detailed description of the methods used to collect and analyze the data will be given.

With the advances in communication technology, spread of information is accelerated where Internet is now seen as another medium for communication. As the access to the Internet got easier with the increasing affordability of mobile devices, online social networks became more of a part of our daily lives. Twitter provides the data from its users openly to non-commercial use without any charge. The twitter social network may contain different kinds of information such as, text, URL addresses, pictures and movies. According to the announcement on the official second quarterly report of Twitter released from their website, the average tweet count per day has reached to 500 million and the average monthly active users reached to 271 million (Twitter Inc., n.d.). This rich content provides a valuable data source for social science studies, as well as marketing and economic researches. However, it is often hard to analyze due to its vast quantity of the data and the complex graph structure it contains.

Twitter uses open source software and provides Application Programming Interfaces (API) to access the features from various platforms. There are three main API's that is used for data collection, REST API that allows gathering information that is relatively stable ,Search API allows to query all the recent available twitter data and Streaming API that allows to mine tweets in real time. In order to use these API's the users should get authantication by registering using OAUTH service. By using these methods Twitter permits to reach almost all data, however with the recent changes on the new version of the Twitter API there are limits on how many requests one can make by using these methods. These timing limits are called "rate limitting" and the rate limit window varies between methods and in API v1.1. rates are divided in 15 minute blocks.

The information can be gathered with the API methods are listed below:

1. **Tweet:** A tweet is the message posted by the user. With a search method we can get the tweets about the queried item. A tweet includes:
   - **Text:** The tweet text that is contained within a tweet.
   - **User:** The user name and id who posted the tweet.
   - **Entities:**
     - **hashtags:** The tags the user posted within a message which can be identified by the '#' symbol.
     - **urls:** Any web addresses the user posted within the tweet.
     - **user mentions:** Any mentions of other users included in the message that can be identified by the '@' symbol.
   - **Created at:** The time the tweet is posted.
   - **Coordinates:** The coordinates of the place the tweet is posted.
   - **Favorite count:** How many times the tweet is favorited.
   - **Retweet count:** How many times the tweet is re-tweeted.
   - **ID:** The string form of identification for the tweet.
   - **In reply to user name:** If the tweet is a reply to a user, the name of the user.
   - **In reply to status id:** If the tweet is a reply to a status, the id of the status.
   - **Language:** The language of the tweet, usually determined by the information the user provides.
   - **Place:** Places mentioned by the user in the tweet.
2. **Friends:** The friendlist of the user, its followees.
3. **Followers:** The followers of the user.
4. **User timelines:** All the tweets the user posted, it is limited to 3,200 recent tweets.
5. **Trending Topics:** The list of the hashtags that are most tweeted about by the users. Trending topics specific to a location may be obtained.

All of the information Twitter provides can be received in JSON documentation format. The follower-followee relation shows the presence of absence of social links between individuals.

As an empirical data, we used the twitter data from users around the world that have commented on the Soma mining disaster in Turkey that took place at 13th of May 2014. Many of the social media users have been moved by the incident and the political debates about the poor working conditions with no safety precautions triggered the protests throughout Turkey. We have gathered time-stamped tweets about the disaster and further looked for another political event, the 1st anniversary of Gezi protests that started in June 2013. The important characteristics of these events is both of them have been the subject of enthusiastic social media topics by two main and opposite sides, which can be discriminated from the hashtags they used while mentioning the events. This property of the topics sometimes were in the form of hashtag wars, where each side tries to move the hashtag up in Trending Topics list, where it will have more chance to be seen by more people. This allowed us to treat the phenomenon as it is similar to the naming game, where users try to reach a consensus on a label to refer to a certain event. The aim to conduct this study is to test our model on an empirical data, and compare our results with the simulation.

In this study we used python-twitter library that is available for Python environment, that can be reached from (https://pypi.python.org/pypi/python-twitter/1.3.1). Twitter REST API and Streaming API v1.1 methods are used to get the data from Twitter.

3.2.1 Statement of Ethical Use

The data used in this study is available to the public use by the website Twitter.com. All the private information gathered from users have been used by securing the anonymity of individuals, and none of the data that may reveal the identity of any user will be published.

3.2.2 Searching for Keywords

Using the search methods provided by the python-twitter library, starting on the day the incident happened (May 13th) we queried every tweet that includes the hashtag 'Soma' inside. We also queried the tweets that uses the trending hashtags about the Soma incident. We have collected tweets with those queries until 21st of May. Later on 31st of May, the anniversary of Gezi protests we also collected tweets that have 'Gezi' in the hashtags or trending hashtags about gezi. We have collected tweets with those hashtags until 4th of June. A total of 113.790 tweets about "Soma" and 25589 tweets about "Gezi" have collected. These tweets included information about:

- user name and user id
- tweet id
- text of the tweet and the entities included within text
- information about when the tweet posted

Tweets about Soma was posted from a total of 46779 different users. Where the tweets about Gezi was posted from 18261 users.

Because we collect only the trending hashtags, we do not claim that we have covered all of the tweets that were about Soma or Gezi. We searched for words in hashtags instead of querying the words in texts to eliminate the possible tweets that contains the words "Soma" or "Gezi", but not actuall about the events we want to track. For example "Gezi" in Turkish is a frequently used word that means "journey", and users that simply talks about their journeys may use this word. Similarly "soma" is also a word used for describing the body of a neuron in biology, not to mention various other uses of the word (Wikipedia, 2014).

3.2.3 Mining User Information

After extracting user names and ids from the collected data, we mined detailed user information for each user. We mined the friends and follower lists of users in "user id" form, and we get the timeline of each user. Users may not allow others to get detailed information about themselves, so we could not get the user data for every user we collected before. We could get the friend-follower information from a total of 46608 users. We also got timeline data from a total of 11009 users, 4269 of them tweeted about both "Soma" and "Gezi".

3.2.4 Working with Data

3.2.4.1 Cleaning the Data

Once we have gathered the relevant data we cleaned it from Twitter bots, which are automated programs that are created to produce tweets according to certain settings. Some of those bots are working as Spam that are mostly used for commercial use, some of them are created to simply promote the popularity of certain topics or users by constantly tweeting and retweeting about them. According to the official announcement of Twitter, 8.5 percent of the Twitter accounts are automated bots (Twitter Inc., 2014). In the paper of Chu et al. (2012), they have listed some methods to find these automated messages in the data:

- Entropy-based approach, using the frequency of the posts that reveals the automation. Some accounts does not have sleep-wake cycle, that they tweet constantly, or some accounts post messages more than one time every minute that may be the indication they are automated bots. Using this method we have cleaned more than 4 percent of our data.
- Spam detection based approach, using the content of the tweets. Some accounts only post advertisements, for example in our data posts texting "Para vermeden beleş takipçi :D" or "Tek tıkla en fazla takipçi veren site" with trending hashtags consisted of almost 4 percent of our data.
- Using account details, automated users tend to have more friends than followers, however celebrities and organizations should be excluded from this list. Moreover, the channel users post their tweets may be another indicator of whether or not a user is a bot. Human users tend to use Twitter on the Web or on mobile applications, where bots and cyborgs mostly use unregistered APIs and RSS feeds.
- Using a decision maker that combines these methods and outputs the decision on whether or not an account is a bot.

We only used the first two method for eliminating the probable bot accounts in our data. We conducted our experiments on both the cleaned data and a sample of the data where we checked the accounts one by one to make sure it contains no automated accounts. 8.2 percent of our initial data was consisting of bots, our initial data consisted of 10202 users.

3.2.4.2 Analyzing the Data

In our analysis, we used the timeline information of every user from our cleaned data and generated two samples from the users according to the dates they posted in order to control the tweeting rates of users and the topics that are being posted. We extracted the users that posted messages with hashtags between the dates "Tue May 13 00:00:00 +0000 2014" and "Mon Jun 02 00:00:00 +0000 2014" for the first sample, "Tue May 13 00:00:00 +0000 2014" and "Sat Jun 13 00:00:00 +0000 2014" for the second sample. Sample-1 consisted of 2071 users, while Sample-2 consisted of 1839 users.

From every user in those samples we extracted all the hashtags about soma, gezi, and all the hashtags during those periods of time. Once we extracted the hashtags about soma h_{soma}, hashtags about gezi h_{gezi} and all hashtags except soma h_{all_s} and all hashtags except gezi h_{all_g} we discarded the outliers in the data in terms of number of hashtags for Sample-1 (M = 24.03, SD = 36.76) and for Sample-2 (M = 28.29, SD = 36.78). After removing the outliers we grouped the users in three subgroups in both of those samples to be able to test our hypothesis in three distinct cases. The subgroups consisted of users who posted a hashtag about soma (h_{soma}), users who posted a hashtag about gezi (h_{gezi}) and the users who posted about both soma and gezi (h_{gs}) by further removing the users that did not post any hashtags about those subjects from those subgroups. The initial data of those subgroups for Sample-1 consisted of N_{soma} = 1563, N_{gezi} = 603, N_{sg} = 595 users; and for Sample-2 N_{soma} = 1445, N_{gezi} = 559, N_{sg} = 551 users.

The hashtags were all converted to lowercase strings, and Turkish ASCII characters are converted to their pairs in English alphabet. Therefore writing 'IsKazasiDegilCinayet' and 'İşKazasıDeğilCinayet' counted as the same.

Next, we calculated the similarity between those hashtags, using cosine similarity metric. Cosine metric measures the angular similarity between two vectors of various dimensions.

$$\cos(\theta) = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{3.10}$$

We calculated the similarities between pairs of users. Therefore for number of n users that posted a hashtag about soma (h_{soma}), there will be $n * (n-1)/2$ number of similarity values for the whole population.

After calculating the similarities between hashtags, we calculated whether there is a correlation between groups of hashtags:

- between hashtags about soma h_{soma} and all hashtags except soma h_{all_s} as r_{soma}
- between hashtags about gezi h_{gezi} and all hashtags except gezi h_{all_g} as r_{gezi}
- between hashtags about soma h_{soma} and hashtags about gezi h_{gezi} as r_{gs}.

Note that the last calculation will show the correlation between two political topics. We calculated the correlation using Pearson's correlation coefficient between the similarity values of every pair of users.

For Sample-1; within the users who posted about soma (N_{soma} = 1559, M = 21.96, SD = 18.31), the correlation between hashtags about soma (h_{soma}) and all hashtags except soma h_{all_s} was significantly and strongly correlated ($r_{soma}(12144520) = .090$, $p < .001$). Within the users who posted about gezi (N_{gezi} = 603, M = 29.70, SD = 19.39), the correlation between hashtags about gezi (h_{gezi}) and all hashtags except gezi h_{all_g} was significantly and strongly correlated ($r_{gezi}(181494) = .096, p < .001$). Within users who posted both about gezi and soma (N_{sg} = 595, M = 30.02, SD = 19.32) the correlation between hashtags about gezi (h_{gezi}) and hashtags about soma h_{soma} was significantly and strongly correlated ($r_{gs}(176708) = .109, p < .001$).

Table 3.1: Correlation Values for Sample-1

	N	M	SD	r	p
soma	1559	21.96	18.31	.090	p < .001
gezi	603	29.70	19.39	.096	p < .001
soma-gezi	595	30.02	19.32	.109	p < .001

For Sample-2 between the users who posted about soma ($N_{soma} = 1444$, $M = 24.62$, $SD = 19.67$), the correlation between hashtags about soma (h_{soma}) and all hashtags except soma h_{all_s} was significantly and strongly correlated ($r_{soma}(1041834) = .088$, $p < .001$). Within the users who posted about gezi ($N_{gezi} = 559$, $M = 26.95$, $SD = 17.95$), the correlation between hashtags about gezi (h_{gezi}) and all hashtags except gezi h_{all_g} was significantly and strongly correlated ($r_{gezi}(155954) = .078$, $p < .001$). Within users who posted both about gezi and soma ($N_{sg} = 551$, $M = 32.90$, $SD = 20.01$) the correlation between hashtags about gezi (h_{gezi}) and hashtags about soma h_{soma} was significantly and strongly correlated ($r_{gs}(151518) = .110, p < .001$).

Table 3.2: Correlation Values for Sample-2

	N	M	SD	r	p
soma	1444	24.62	19.67	.088	p<.001
gezi	559	26.95	17.95	.078	p < .001
soma-gezi	551	32.90	20.01	.110	p < .001

3.2.4.3 Creating a Sub-Sample

Although we cleaned the data using aforementioned methods in order to make sure our data does not include any bots, we created a sub-sample from the previously cleaned data that is cleaned by hand. We analysed this data as the previous one and got the results below.

We extracted the users from our subsample that posted messages with hashtags between the dates "Tue May 13 00:00:00 +0000 2014" and "Mon Jun 02 00:00:00 +0000 2014" for the first sub-sample, "Wed May 13 00:00:00 +0000 2014" and "Sat Jun 13 00:00:00 +0000 2014" for the second sub-sample. Subsample-1 consisted of 209 user data, while Subsample-2 consisted of 190 user data. We eliminated the outliers in the data in terms of number of hashtags for Subsample-1 ($M = 46.00$, $SD = 41.44$) and for Subsample-2 ($M = 52.83$, $SD = 40.14$).

After calculating the cosine similarity between users as described above, we computed the Pearson correlation coefficient between the similarity values of r_{soma}, r_{gezi} and r_{gs} for every pair of users.

For Subsample-1; within the users who posted about soma ($N_{soma} = 201$, $M = 38.64$, $SD = 25.11$), the correlation between hashtags about soma (h_{soma}) and all hashtags except soma h_{all_s} was not significantly correlated ($r_{soma}(20100) = .010$, $p = .168$). Within the users who posted about gezi ($N_{gezi} = 184, M = 41.29, SD = 26.42$), the correlation between hashtags about gezi (h_{gezi}) and all hashtags except gezi h_{all_g} was not significantly correlated ($r_{gezi}(16836) = .011$, $p = .167$). Within users who posted both about gezi and soma ($N_{sg} = 183$, $M = 40.81$, $SD = 25.70$) the correlation between hashtags about gezi (h_{gezi}) and hashtags about soma h_{soma}

was significantly and strongly negatively correlated ($r_{gs}(16653) = -.037, p < .001$).

Table 3.3: Correlation Values for Subsample-1

	N	M	SD	r	p
soma	201	38.64	25.11	.010	p = .168
gezi	184	41.29	26.42	.011	p = .167
soma-gezi	183	40.81	25.70	-.037	p < .001

For Subsample-2 between the users who posted about soma ($N_{soma} = 181$, $M = 45.73$, $SD = 28.48$), the correlation between hashtags about soma (h_{soma}) and all hashtags except soma h_{all_s} was not significantly correlated ($r_{soma}(16290) = -.012$, $p = .118$). Within the users who posted about gezi ($N_{gezi} = 169$, $M = 46.82$, $SD = 23.18$), the correlation between hashtags about gezi (h_{gezi}) and all hashtags except gezi h_{all_g} was not significantly correlated ($r_{gezi}(14196) = -.005$, $p = .57$). Within users who posted both about gezi and soma ($N_{sg} = 169$, $M = 46.82$, $SD = 23.18$) the correlation between hashtags about gezi (h_{gezi}) and hashtags about soma h_{soma} was significantly and negatively strongly correlated ($r_{gs}(14196) = -.034, p < .001$).

Table 3.4: Correlation Values for Subsample-2

	N	M	SD	r	p
soma	181	45.73	28.48	-.012	p = .118
gezi	169	46.82	23.18	-.005	p = .57
soma-gezi	169	46.82	23.18	-.034	p < .001

CHAPTER 4

CONCLUSION AND DISCUSSION

Interactions between individuals drives cultural traits as a process of meaning making and the outcome of interaction determines the behavior of individuals. Shared experiences during these interactions allow groups of individuals to reach a consensus on their linguistic behavior: language. Culture and language as being the direct result of interpersonal interactions, shapes how later interactions will take place. As a result, culture and language is densely connected. This study is aimed to show the effect of culture in language as a communication medium, with the hypothesis that we can directly observe the effects of cultural traits in linguistic behavior and predict the behavior of individuals by using their linguistic history. This assumption is modeled using multi-agent simulation approach where a population of agents are represented in a directed network and the communication between agents is in the form of a naming game. The cultural selection forces both in selecting a communication partner and a communication topic is implemented via functions that operates on the internal structures of agents and communicative interactions between agents.

The model showed that the selection of the communication partner results in a topology formation in a homogeneous initial state and selection of the topic drives the population in a multilingual state where some agents do not prefer to talk about some topics. These preferences further determines the cultural subgroups. As a result of topic preference, the population almost never reaches a convergent state, however the rate of successful communications always inclines due to partner selection according to previous successful interactions. Therefore, the agents in the population have to use more memory than classical naming games. When this model tested for the argued hypothesis, a strong correlation have been found between the linguistic behavior of the agents and their lexicons. Some of the simulations showed no correlation or a weak correlation between these traits between the agents who avoid conflicting communications by chance.

After implementing the model in the simulation environment, the hypothesis further tested on a real life social network in a case that highly resembles naming game dynamics. A popular political topic is selected in the online social microblogging website, Twitter, as the source of experimental data, where the hashtags used to refer to that topic are collected for every user. This data is further used to predict the behavior of the same users on another popular topic. The topics was deliberately selected about popular political subjects that caused the users to agree on a hashtag about the topics. This allowed us to form a relation between the naming game where the agents try to reach a consensus about a particular topic. The analyses on the data showed a strong correlation between the similarities of previous usage of hashtags of agents and the hashtags about a particular subject, allowing us to conclude that we may predict the behavior of a user on a particular topic by looking at similar users' posting behavior. Therefore, we may conclude that the suggested model was successful at predicting the behavior of users looking at previous data. However, the subsample we used in the study did not show a correlation except for the negative correlation between the topics "gezi" and "soma". This conflicting result may be due to the small sample size that failed to represent the population. In order to reach a precise conclusion, a cross correlation between a number of subsamples should be used. Due to time limitations this solution could not be performed, but postponed to further studies.

Moreover, the correlation between the topics "gezi" and "soma" was expected to be stronger, however Pearson's correlation coefficient will not account for how strong the correlation is. Nevertheless, bearing in mind that correlation does not necessarily mean causation, we can not directly conclude that this result shows a one-to-one correspondence to the suggested model.

The primary limitation of this study was the amount of time and memory necessary to carry on the computations. Some simulations took more than two days to compute and analyze and had to be repeated for each parameter setting for a couple of times in order to use the average values. Therefore, a full coverage of the number of agents, objects and parameter sets could not be examined in the scope of this study. It should also be noted that the aim of this study is limited to the signal-meaning pairs for sake of simplicity. Therefore, there is no connection between the objects in the environment and objects does not have any properties of their own. Moreover, the agents are not able to connect the words in their lexicons to each other. Consequently, the linguistic behavior of the agents is primitive in the sense that it bears no syntactic element or connection.

Additionally, the amount of time to gather the empirical data from Twitter was more than two months, by using five computers due to "rate limiting" of Twitter servers. Consequently, the methods used in this study are not suitable for real-time classification of users even though the accuracy of determining similar users is high. Additionally, information of some users could not be mined due to user preferences that would not permit sharing of personal information. Also, full access to all the timeline information for every user was not granted by Twitter, resulting in dismissing some of the user information due to absent data. This caused the elimination of about eighty percent of the users, most of which are active users of Twitter.

For future studies, we may gather more data that covers the whole timeline information of every Twitter user. This will allow including the information of the are more active users to the analysis, therefore having a more complete dataset. Additionally, for analysis we may try to compare all the hashtags instead of only focusing a determined topic, or select a neutral topic that involves no conflict between users. Furthermore the text data may be used for additional sentiment analysis to see the emotional value one attaches to the particular topic and analyze the similarities between users.

Future work for the simulation should first include a more comprehensive examination of parameter settings. In this study, the maximum value γ_{max} for topic list, minimum and maximum values (δ_{max}, δ_{min}) for the network links between agents was not examined. In addition to those values, a wider parameter space for already examined parameters may be used and compared with previous results.

A curious extension to the current simulation would be introducing a new object to the environment in a population that already interacted about other objects for some time. This simple addition to the simulation will create a more realistic situation that resembles the Twitter case where new topics always emerges in the population. It will allow the examination of the behavior of the agents where there is a new topic to talk about.

Another valuable addition to both the simulation and empirical study is to expand the hypothesis by using incoming information from connected pairs. Although our hypothesis set out from the idea of spread of information and meaning creation, it does not take into account the information an individual gets from the environment from its connected pairs. In the Twitter case, a user will constantly get information from its friends that effects its future behavior. We may examine this effect via controlling the posts of the friends of each user. This effect may be simulated in a setting where an agent could get information about every utterance in its network. This version of the naming game would require an extensive analysis of the topology, where network theory will have a higher significance.

Bibliography

Aggarwal, C. C. (2011). *An introduction to social network data analytics*. Springer.

Axelrod, R. (1997). The dissemination of culture a model with local convergence and global polarization. *Journal of conflict resolution 41*(2), 203–226.

Baronchelli, A., L. Dall'Asta, A. Barrat, and V. Loreto (2006). Topology-induced coarsening in language games. *Physical Review E 73*(1), 015102.

Baronchelli, A., R. Ferrer-i Cancho, R. Pastor-Satorras, N. Chater, and M. H. Christiansen (2013). Networks in cognitive science. *Trends in cognitive sciences 17*(7), 348–360.

Baronchelli, A., V. Loreto, and L. Steels (2008). In-depth analysis of the naming game dynamics: the homogeneous mixing case. *International Journal of Modern Physics C 19*(05), 785–812.

Barrat, A., A. Baronchelli, L. Dall'Asta, and V. Loreto (2007). Agreement dynamics on interaction networks with diverse topologies. *Chaos: An Interdisciplinary Journal of Nonlinear Science 17*(2), 026111.

Bellomo, N. and C. Dogbe (2008). On the modelling crowd dynamics from scaling to hyperbolic macroscopic models. *Mathematical Models and Methods in Applied Sciences 18*(supp01), 1317–1345.

Bender, A., E. Hutchins, and D. Medin (2010). Anthropology in cognitive science. *Topics in Cognitive Science 2*(3), 374–385.

Berlo, D. K. (1960). *The process of communication: An introduction to theory and practice*. Holt, Rinehart and Winston New York.

Cangelosi, A. and D. Parisi (2002). *Simulating the evolution of language*, Volume 1. Springer London.

Castellano, C., S. Fortunato, and V. Loreto (2009). Statistical physics of social dynamics. *Reviews of modern physics 81*(2), 591.

Christiansen, M. H. and S. Kirby (2003). Language evolution: Consensus and controversies. *Trends in cognitive sciences 7*(7), 300–307.

Chu, Z., S. Gianvecchio, H. Wang, and S. Jajodia (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *Dependable and Secure Computing, IEEE Transactions on 9*(6), 811–824.

Craig, R. T. (1999). Communication theory as a field. *Communication theory 9*(2), 119–161.

Croft, W. (2008). Evolutionary linguistics. *Annual Review of Anthropology 37*(1), 219.

Dall'Asta, L., A. Baronchelli, A. Barrat, and V. Loreto (2006). Nonequilibrium dynamics of language games on complex networks. *Physical Review E 74*(3), 036105.

Epstein, J. M. (2002). Modeling civil violence: An agent-based computational approach. *Proceedings of the National Academy of Sciences of the United States of America 99*(Suppl 3), 7243–7250.

Galam, S., Y. Gefen, and Y. Shapir (1982). Sociophysics: A new approach of sociological collective behaviour. i. mean-behaviour description of a strike. *Journal of Mathematical Sociology 9*(1), 1–13.

Girvan, M. and M. E. Newman (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences 99*(12), 7821–7826.

Hegselmann, R. and U. Krause (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation 5*(3).

Helbing, D. (2001). Traffic and related self-driven many-particle systems. *Reviews of modern physics 73*(4), 1067.

Huston, T. L. and G. Levinger (1978). Interpersonal attraction and relationships. *Annual review of psychology 29*(1), 115–156.

Hutchins, E. and C. M. Johnson (2009). Modeling the emergence of language as an embodied collective cognitive activity. *Topics in Cognitive Science 1*(3), 523–546.

Kempe, D., J. Kleinberg, S. Oren, and A. Slivkins (2013). Selection and influence in cultural dynamics. *arXiv preprint arXiv:1304.7468*.

Kempe, D., J. Kleinberg, and É. Tardos (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146. ACM.

Kirby, S. (2002). Natural language from artificial life. *Artificial life 8*(2), 185–215.

Klemm, K., V. M. Eguíluz, R. Toral, and M. San Miguel (2003). Global culture: A noise-induced transition in finite systems. *Physical Review E 67*(4), 045101.

Lasswell, H. D. (1948). The structure and function of communication in society. *The communication of ideas 37*.

Lipowska, D. and A. Lipowski (2012). Naming game on adaptive weighted networks. *Artificial life 18*(3), 311–323.

Loreto, V., A. Baronchelli, A. Mukherjee, A. Puglisi, and F. Tria (2011). Statistical physics of language dynamics. *Journal of Statistical Mechanics: Theory and Experiment 2011*(04), P04006.

Lu, Q., G. Korniss, and B. K. Szymanski (2009). The naming game in social networks: community formation and consensus engineering. *Journal of Economic Interaction and Coordination 4*(2), 221–235.

Lull, J. (2002). *Culture in the communication age*. Routledge.

Macal, C. M. and M. J. North (2008). Agent-based modeling and simulation: Abms examples. In *Proceedings of the 40th Conference on Winter Simulation*, pp. 101–112. Winter Simulation Conference.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 415–444.

Milgram, S. (1967). The small world problem. *Psychology today 2*(1), 60–67.

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

Newman, M. E. and J. Park (2003). Why social networks are different from other types of networks. *Physical Review E 68*(3), 036122.

Schramm, W. (1954). How communication works. *The process and effects of mass communication*, 3–26.

Senghas, A., S. Kita, and A. Özyürek (2004). Children creating core properties of language: Evidence from an emerging sign language in nicaragua. *Science 305*(5691), 1779–1782.

Shannon, C. (1948, July, October). *Bell System Technical Journal 27*, 379–423, 623–656.

Shannon, C. E. and W. Weaver (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Singla, P. and M. Richardson (2008). Yes, there is a correlation:-from social networks to personal behavior on the web. In *Proceedings of the 17th international conference on World Wide Web*, pp. 655–664. ACM.

Steels, L. (2000). Language as a complex adaptive system. *Proceedings of the 6th International Conference on Parallel Problem Solving from Nature*.

Steels, L. (2011). Modeling the cultural evolution of language. *Physics of Life Reviews 8*(4), 339–356.

Steels, L. (2012). Self-organization and selection in cultural language evolution. *Experiments in Cultural Language Evolution. John Benjamins, Amsterdam.*

Steels, L., T. Belpaeme, et al. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and brain sciences 28*(4), 469–488.

Sun, J. and J. Tang (2011). A survey of models and algorithms for social influence analysis. In *Social Network Data Analytics*, pp. 177–214. Springer.

Vygotsky, L. S. (1997). *The collected works of LS Vygotsky: Problems of the theory and history of psychology*, Volume 3. Springer.

Vylder, B. D. and K. Tuyls (2006, October). How to reach linguistic consensus: A proof of convergence for the naming game. *Journal of Theoretical Biology 242*(4), 818–831.

Wagner, K., J. A. Reggia, J. Uriagereka, and G. S. Wilkinson (2003). Progress in the simulation of emergent communication and language. *Adaptive Behavior 11*(1), 37–69.

Watts, D. J. (2004). The" new" science of networks. *Annual review of sociology*, 243–270.

Wittgenstein, L. (1953). *Philosophical investigations. Philosophische Untersuchungen.* Macmillan.

APPENDIX A

CORRELATION VALUES FOR THE SIMULATION

Table A.1: Correlation values for different numbers of objects and population $\theta_{success} = 0.4$, $\theta_{failure} = 0.2$ in the case of a naming game with topic selection and pair selection mechanisms.

Number of Agents	Number of Objects	ω_{min}	$r_{lexicon-interest}$	$p_{lexicon-interest}$
10	10	0	0.0584033658	0.7031420638
20	10	0	0.8898433531	5.59E-066
30	10	0	0.2885485268	3.17E-009
40	10	0	0.1941181046	4.63E-008
50	10	0	0.3798820727	1.13E-041
60	10	0	0.1374750847	2.00E-008
70	10	0	0.1864494428	2.89E-019
80	10	0	0.0628502055	0.0006698376
90	10	0	0.186338881	6.26E-032
100	10	0	0.2246007958	1.22E-057
10	20	0	0.8276206168	2.37E-012
20	20	0	0.750778828	1.07E-035
30	20	0	0.722094447	2.72E-071
40	20	0	0.702196909	-6.77E-117
50	20	0	0.7583706328	1.44E-229
60	20	0	0.6356677544	5.56E-201
70	20	0	0.3960481764	1.63E-091
80	20	0	0.5285162747	6.24E-227
90	20	0	0.310130088	5.08E-090
100	20	0	0.4236722928	7.24E-215

Number of Agents	Number of Objects	ω_{min}	$r_{lexicon-interest}$	$p_{lexicon-interest}$
10	5	0	0.6518105724	1.23E-006
20	5	0	0.4831858488	1.65E-012
30	5	0	-0.1511337262	0.0015709307
40	5	0	-0.2045695056	8.15E-009
50	5	0	-0.0539417837	0.0591053023
60	5	0	0.0258854923	0.2763960119
70	5	0	0.0015851481	0.9379409318
80	5	0	-0.0518237651	0.0035680669
90	5	0	0.022362408	0.1570855896
100	5	0	-0.0098844081	0.48688542
10	5	0.1	0.6069281136	9.87E-006
20	5	0.1	-0.0690285389	0.3439735322
30	5	0.1	0.2198943381	3.66E-006
40	5	0.1	-0.0923599529	0.00985561
50	5	0.1	-0.0677829288	0.0176581331
60	5	0.1	0.0081578256	0.7316179214
70	5	0.1	-0.0846115551	3.14E-005
80	5	0.1	-0.068854859	0.0001072108
90	5	0.1	-0.0739830595	2.77E-006
100	5	0.1	0.0200773895	0.1578454131
10	10	0.1	0.4998704523	4.72E-004
20	10	0.1	-0.2468266589	0.0005967385
30	10	0.1	-0.1260009748	8.52E-003
40	10	0.1	0.136969717	0.0001243842
50	10	0.1	0.0659906242	0.020897457
60	10	0.1	-0.1102704364	3.32E-006
70	10	0.1	-0.1292410712	1.83E-010
80	10	0.1	-0.0651378538	0.0002482155
90	10	0.1	-0.0050011641	7.52E-001
100	10	0.1	-0.0443464342	0.0018036325
10	20	0.1	0.4721180956	0.0010592314
20	20	0.1	-0.1332311549	0.0668733632
30	20	0.1	-0.0063213663	0.8954078663
40	20	0.1	-0.1365548289	0.0001304611
50	20	0.1	-0.078070177	0.0062600449
60	20	0.1	0.0068374613	0.773758983
70	20	0.1	-0.086404154	2.12E-005
80	20	0.1	-0.0315635566	0.0760548965
90	20	0.1	-0.056245407	0.0003691602
100	20	0.1	-0.070868793	6.00E-007

Table A.2: Correlation between the similarity of utterances and lexicons of agents for different numbers of objects and population $\theta_{success} = 0.4$, $\theta_{failure} = 0.2$ in the case of a naming game with topic selection and pair selection mechanisms.

Number of Agents	**Number of Objects**	ω_{min}	$r_{lexicon-utterances}$	$p_{lexicon-utterances}$
10	10	0	0.8077789882	2.00E-011
20	10	0	0.5467039145	3.35E-016
30	10	0	0.5945000923	6.39E-043
40	10	0	-0.0421571961	0.2395864745
50	10	0	-0.0124844684	0.6624541796
60	10	0	-0.0010365047	0.9652420825
70	10	0	-0.0338558046	0.0962365148
80	10	0	0.144288295	3.62E-016
90	10	0	-0.0292972715	0.0637536077
100	10	0	-0.0210989229	0.1377476197
10	20	0	0.9413863801	0
20	20	0	0.9186169551	0
30	20	0	0.9043466791	0
40	20	0	0.8680308662	0
50	20	0	0.8934495088	0
60	20	0	0.8839387765	0
70	20	0	0.8566126311	0
80	20	0	0.7707200696	0
90	20	0	0.5350404786	0
100	20	0	0.5172088884	0
10	5	0	0.6438119693	1.83E-006
20	5	0	0.4239907742	1.09E-009
30	5	0	-0.0139691277	0.7714145397
40	5	0	-0.0916360364	0.010450965
50	5	0	-0.06621987	0.0204561815
60	5	0	0.1370783401	7.02E-009
70	5	0	-0.0045339037	0.8237738014
80	5	0	-0.1055778013	2.69E-009
90	5	0	0.0227687385	0.1496812826
100	5	0	0.0210036747	0.1395338187

Table A.3: Correlation between the similarity of utterances and lexicons of agents for different numbers of objects and population $\theta_{success} = 0.4$, $\theta_{failure} = 0.2$ in the case of a naming game with only topic selection mechanism with no network.

Number of Agents	Number of Objects	$r_{lexicon-utterances}$	$p_{lexicon-utterances}$
10	5	0.0813919608	0.595063524
20	5	-0.0254326933	0.7276085313
30	5	0.0928982256	0.0528481386
40	5	0.0928982256	0.0528481386
50	5	-0.0722371522	0.0114380297
60	5	-0.0691719656	0.0035959291
70	5	-0.0714235241	0.0004437273
80	5	-0.047954122	0.0070141498
90	5	-0.0062797603	0.6911499884
100	5	0.0858469898	1.45E-009
10	10	0.007190758	0.9626086805
20	10	0.0565986693	0.437965946
30	10	0.0953764108	0.046806927
40	10	-0.1423418161	6.63E-005
50	10	-0.0342687927	0.2307086882
60	10	0.1855979871	3.51E-015
70	10	-0.0171592513	0.3992974872
80	10	0.0420352398	0.0181240818
90	10	-0.066684233	2.41E-005
100	10	0.1652037073	1.26E-031
10	20	0.1088493579	0.4766163356
20	20	-0.2272632951	0.0016143496
30	20	0.0903649938	0.0596810378
40	20	0.1077716742	0.0025795192
50	20	-0.0436640191	0.1266593447
60	20	-0.0530309342	0.0256759777
70	20	0.0941907681	3.54E-006
80	20	0.0327177352	0.0659217485
90	20	-0.0622397436	8.10E-005
100	20	0.0804897805	1.42E-008

Table A.4: Correlation values for different numbers of objects and population $\theta_{success} = 0.4$, $\theta_{failure} = 0.2$ in the case of a naming game with only pair selection mechanism in a network with $\omega_{min} = 0.1$.

Number of Agents	**Number of Objects**	ω_{min}	$r_{lexicon-interest}$	$p_{lexicon-interest}$
10	5	0.1	0.0919347637	0.5480818909
20	5	0.1	0.173031894	0.0169690735
30	5	0.1	0.0221433933	0.6451109546
40	5	0.1	-0.0260750666	0.4671079093
50	5	0.1	0.0743188609	0.0092653821
60	5	0.1	0.0580178989	0.0146375532
70	5	0.1	-0.0436342356	0.0320158227
80	5	0.1	0.0522618891	0.0032960325
90	5	0.1	-0.0152809527	0.3336386635
100	5	0.1	-0.0421496397	0.0030164248

Table A.6: Correlation values between the words spoken for each object and the rest of the lexicon in different population size $\theta_{success} = 0.4$, $\theta_{failure} = 0.2$ and $\omega_{min} = 0$ in the case of a naming game with topic selection and pair selection mechanisms.

Number of Agents	**Object Number**	$r_{lexicon-interest}$	$p_{lexicon-interest}$
10	1	0.5313	0.0001
	2	-0.1006	0.5109
	3	0.2865	0.0563
	4	0.1413	0.3545
	5	0.4724	0.0010
20	1	0.2209	0.0022
	2	0.1336	0.0660
	3	0.1645	0.0233
	4	-0.0485	0.5064
	5	-0.0416	0.5684
30	1	0.3241	4.2369e-12
	2	-0.1941	4.5879e-05
	3	0.0712	0.1379
	4	0.1100	0.0217
	5	-0.0528	0.2717
40	1	-0.0251	0.4846
	2	0.1890	1.0459e-07
	3	0.1650	3.5915e-06
	4	0.0478	0.1825
	5	0.1473	3.6236e-05

Table A.5: Correlation values for different numbers of objects and population $\theta_{success} = 0.4$, $\theta_{failure} = 0.2$ in the case of a naming game with only topic selection mechanism with no network.

Number of Agents	**Number of Objects**	$r_{lexicon-interest}$	$p_{lexicon-interest}$
10	5	0.0813919608	0.595063524
20	5	-0.0254326933	0.7276085313
30	5	0.0928982256	0.0528481386
40	5	0.0928982256	0.0528481386
50	5	-0.0722371522	0.0114380297
60	5	-0.0691719656	0.0035959291
70	5	-0.0714235241	0.0004437273
80	5	-0.047954122	0.0070141498
90	5	-0.0062797603	0.6911499884
100	5	0.0858469898	1.45E-009
10	10	0.007190758	0.9626086805
20	10	0.0565986693	0.437965946
30	10	0.0953764108	0.046806927
40	10	-0.1423418161	6.63E-005
50	10	-0.0342687927	0.2307086882
60	10	0.1855979871	3.51E-015
70	10	-0.0171592513	0.3992974872
80	10	0.0420352398	0.0181240818
90	10	-0.066684233	2.41E-005
100	10	0.1652037073	1.26E-031
10	20	0.1088493579	0.4766163356
20	20	-0.2272632951	0.0016143496
30	20	0.0903649938	0.0596810378
40	20	0.1077716742	0.0025795192
50	20	-0.0436640191	0.1266593447
60	20	-0.0530309342	0.0256759777
70	20	0.0941907681	3.54E-006
80	20	0.0327177352	0.0659217485
90	20	-0.0622397436	8.10E-005
100	20	0.0804897805	1.42E-008

Number of Agents	**Object Number**	$r_{lexicon-interest}$	$p_{lexicon-interest}$
50	1	-0.0319	0.2638
	2	0.1410	7.1409e-07
	3	0.1184	3.2279e-05
	4	0.0525	0.0660
	5	0.0222	0.4375
60	1	-0.0785	0.0009
	2	0.0684	0.0039
	3	0.1076	5.6516e-06
	4	0.1326	2.1330e-08
	5	0.0409	0.0856
70	1	-0.0230	0.2570
	2	0.0983	1.2776e-06
	3	0.0284	0.1625
	4	-0.0509	0.0123
	5	-0.0486	0.0169
80	1	-0.0728	4.222e-05
	2	-0.0484	0.0064
	3	0.2175	3.8687e-35
	4	-0.0159	0.3711
	5	-0.0731	3.9173e-05
90	1	-0.0277	0.0833
	2	-0.0016	0.9170
	3	-0.0026	0.8657
	4	-0.0376	0.0174
	5	0.0107	0.4993
100	1	0.0822	7.0097e-09
	2	-0.0180	0.2048
	3	-0.0066	0.6428
	4	-0.0838	3.5683e-09
	5	0.0766	6.7497e-08

TEZ FOTOKOPİ İZİN FORMU

ENSTİTÜ

Fen Bilimleri Enstitüsü	
Sosyal Bilimler Enstitüsü	
Uygulamalı Matematik Enstitüsü	
Enformatik Enstitüsü	X
Deniz Bilimleri Enstitüsü	

YAZARIN

Soyadı : YALÇIN
Adı : ÖZGE NİLAY
Bölümü : BİLİŞSEL BİLİMLER

TEZİN ADI (İngilizce) : MODELING AND PREDICTING THE EFFECT OF CULTURE IN COMMUNICATION: A MIXED STUDY USING NAMING GAME AND SOCIAL NETWORKS

TEZİN TÜRÜ : Yüksek Lisans X Doktora

1. Tezimin tamamı dünya çapında erişime açılsın ve kaynak gösterilmek şartıyla tezimin bir kısmı veya tamamının fotokopisi alınsın. X
2. Tezimin tamamı yalnızca Orta Doğu Teknik Üniversitesi kullancılarının erişimine açılsın. (Bu seçenekle tezinizin fotokopisi ya da elektronik kopyası Kütüphane aracılığı ile ODTÜ dışına dağıtılmayacaktır.)
3. Tezim bir (1) yıl süreyle erişime kapalı olsun. (Bu seçenekle tezinizin fotokopisi ya da elektronik kopyası Kütüphane aracılığı ile ODTÜ dışına dağıtılmayacaktır.)

Yazarın imzası .... Tarih ....