Intimate connections : an agent-based model of the relationship between social network structure and language typology

Lou-Magnuson, Matthew Elliot

2018

Lou-Magnuson, M. E. (2018). Intimate connections : an agent-based model of the relationship between social network structure and language typology. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/82953>

<https://doi.org/10.32657/10220/46659>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

**Intimate Connections: An agent-based
model of the relationship between social
network structure and language typology.**

Matthew Lou-Magnuson

SCHOOL OF HUMANITIES

2018

**Intimate Connections: An agent-based
model of the relationship between
social network structure and language
typology.**

Matthew Lou-Magnuson

School of Humanities

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2018

Intimate Connections: An agent-based model
of the relationship between social network
structure and language typology.

Matthew Elliot Lou-Magnuson

School of Humanities

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

2018

November 14, 2018

Acknowledgements

Without hesitation, two people are directly responsible for my completion of this work, my advisor, Dr. Luca Onnis, and my wife, Shirley. But for the continued support and encouragement of these two individuals, I would have never completed this dissertation. Thank you both.

In addition, I would like to thank my MA supervisor, Dr. Leon Serafim, for first encouraging me to make the academic move from liturgical studies into the world of historical linguistics, and my BA advisor, Takemoto-Sensei, for sparking an interest in the alternative views that culture and language can provide. For their long-term investment in my desire to go to school and go all the way, I thank my parents, for accepting a son-in-law who has been in school since she met him 11 years ago, my mother-in-law, Liana, and finally for her emotional support, my cat Hua Hua.

Finally, I would like to thank the untold generations of Lushootseed speakers of Puget Sound. The existence of their complex and beautiful language engendered in me the wonder at the diversity of what language can encompass.

Any errors and omissions hereafter are of my own fault and should not reflect on those mentioned above.

Summary

A long standing question concerning natural human language is, why are some languages more complex than others? More specifically, why do some languages utilize intricate morphological structure – in some entire sentences often consist of a single complex word – while in others, all words are void of internal structure, and rely only on ordering and contextual clues to signal grammatical dependencies. There is no intrinsic reason for one strategy over the other, and attested human language spans the spectrum of possibilities.

On the one hand, linguists have proposed a number of theories for why this diversity exists, most centering on the intimate social structure of the speakers as the explanatory factor for increases in morphological composition, and the addition of second-language speakers as the explanatory factor for decreases in morphological composition. However, this work is based on inductive observation of extant language, and at present descriptive and untestable. Furthermore, the time span over which such morphological developments would take place is at least an order of magnitude greater than that reachable with current data and reconstruction methods.

On the other hand, much work in the hard sciences (particularly cosmology and astrophysics) has overcome similar time gaps using computational simulations. Indeed, sound mathematical models of social networks, and complex networks in general, have emerged and been applied to linguistic issues such as the diffusion of lexical innovation. However, the level of representation in such models is at a far more abstract level than that used in linguistics, and even more problematic, the mechanisms of learning and change employed in these models often make patently unrealistic assumptions, for example, language spreads like a virus.

In this thesis, an initial attempt is made to bridge this gap between ideas and means. First, two variables of interest are identified to quantify the notions of compositional complexity and intimacy - synthesis and the global clustering coefficient, respectively. Synthesis is the average proportion of bound morphemes per word in a representative corpus, and was developed specifically to quantify the degree of morphological composition used in a

language. The global clustering coefficient measures the proportion of triangular connections in a complex network, and these triangular connections have been directly connected in the literature to the notion of intimacy.

To explore the relationship between these variables, two novel computational models are created, and are to my knowledge the first to incorporate the fundamental observations within linguistics concerning the development of bound morphology. Specifically, both models have distinct diffusion and intergenerational transmission stages. In addition, the learning mechanisms are differentiated to account for the difference between the learning of novel lexical forms through diffusion, and the slow, incremental grammatical changes that accrue through language transmission. More specifically, the three subprocesses of extension, reanalysis, and repetition identified in the linguistics literature as driving the development of bound morphology are independently addressed.

First, a high-level model that is computationally cheap is developed. It models the capacity of various social networks to support the repeated application of these processes responsible for creating bound morphology. However, this level of abstraction lacks the structure required to measure synthesis directly, and so a low-level model in which the processes are directly modeled on a symbolic meaning-signal structure is also presented.

The results of these two models are in agreement, and find that the physical network measure of the global clustering coefficient predicts support for the processes responsible for bound morphology in the high-level model, as well as predicting higher levels of synthesis in the low-level model. In addition, by testing realistic human social network topologies, it was determined that hub agents that emerge as social networks grow in size, counteract this effect of clustering by serving to reshape the input and interrupt the continuation of the transmission process. This result provides both a mechanistic account of the observed correlation of social structure on increases in synthesis, as well as an alternative explanation for decreases in synthesis that do not involve second-language speakers.

Beyond these initial insights, this thesis demonstrates that the computational study of typological change helps extend our knowledge beyond what can be naturally observed. In this case, it provides inroads to understanding how the intimacy of small societies mechanistically relates to the observed correlations seen in the typological distribution of synthesis.

List of Figures

1.1	Summary of Simulation 1 and Simulation 2.	6
1.2	Summary of Simulation 3, 4, and 5.	7
2.1	Summary of Results from Simulation 1 and Simulation 2. . .	35
2.2	Summary of Results from Simulation 3, 4, and 5.	36
2.3	Simulation 1, 25-agent network results: the total average proportion of innovators is displayed on the y axis, and the communication round number as a measure of time on the x axis.	44
2.4	Simulation 1, 125-agent network results: the total average proportion of innovators is displayed on the y axis, and the communication round number as a measure of time on the x axis.	45
2.5	Simulation 2, 25-agent network results: the total average proportion of innovators displayed on the y axis, and the communication round number as a measure of time on the x axis.	49
2.6	Simulation2, 125-agent network results: the total average proportion of innovators displayed on the y axis, and the communication round number as a measure of time on the x axis.	50
2.7	Schematic of Horizontal Communication Procedure.	61
2.8	Schematic of Compatible Signals	62
2.9	Schematic of Compatible Signals	65
2.10	Example of Random networks used in Simulation 1, with connection probability 0.2 (top) and connection probability 0.7 (bottom). Model parameters are listed to the right.	70
2.11	Results of Simulation 3: mean levels of linguistic reanalysis as a function of network connectivity over 1000 generations.	73
2.12	Local Transitivity: mean levels of linguistic reanalysis as a function of local transitivity. The top two graphs and bottom left are averaged over the individual generation listed, and the bottom right is averaged over all generations. The blue line is the LOESS fit provide by the ggplot2 package in R	74

2.13	Distribution of Local Transitivity: agent counts are taken from all networks generated at the listed connection probability.	76
2.14	The Complete (top left) and Hierarchical (bottom left) networks used in Simulation 4 exemplify connectivity properties of two common types of human social communities respectively: the village and the modern city. The transitivity measures for all networks (BA and Random not pictured) are given to the right (the Complete network depicts only 10 agents for clarity).	79
2.15	Simulation 4, 25-agent network results: mean levels of linguistic reanalysis in four different 25-agent network architectures for 1000 generations. 100 replications per network type.	82
2.16	Simulation 5, 125-agent network results: mean levels of linguistic reanalysis in two different 125-agent network architectures for 500 generations. 100 replications per network type.	82
2.17	Study of Hierarchical Network 3: top figure displays average reanalysis levels across all 100 generations of the network. The middle and bottom figures center around the large drop in reanalysis at generation 500, and are broken down into the physical clusters of agents (middle) and central hub agent vs. others (bottom).	84
2.18	Interplay of Parameters. On the x-axis are the number of rounds of communication allowed before transmission, and the color codes for values of the variation probability, the multiplier used to determine when innovations lead to reanalyses. The y-axis is the mean number of generations between collapses.	88
2.19	Simulation 5, 125-agent random network results: mean levels of linguistic reanalysis as a function of network connectivity over 500 generations.	93
2.20	Simulation 5: interaction effect of topology x size for Complete and Hierarchical.	93
2.21	Simulation 5, 125-agent random network results: mean levels of linguistic reanalysis as a function of network connectivity over 500 generations.	95
3.1	Schematic of Compatible Signals	107
3.2	Example of a Search Frame. For simplicity, only the signal side of the search frame is shown above. In the model an analogous representation of the meaning space would also be present.	113

3.3	Abstract & Generalize	122
3.4	Simplified English ARS	132
3.5	Morphological Revision	133
3.6	Simplified English ARS (Including Morphological Revision) .	134
3.7	Complete Simplified Lushootseed ARS	136
3.8	English Rewrite Dependency Graph	141
3.9	Lushootseed Rewrite Dependency Graph	144
3.10	Results of Simulation 6, 25-Agent Networks: On the x-axis is the generation number as measure of time, and the average synthesis of the language on the y-axis. Shaded regions are the LOESS estimates supplied by the ggplot2 package in R.	151
3.11	Results of Simulation 6, 125-Agent Networks: On the x-axis is the generation number as measure of time, and the average synthesis of the language on the y-axis. Shaded regions are the LOESS estimates supplied by the ggplot2 package in R.	151
3.12	Results of Simulation 7, 25-Agent Networks: On the x-axis is the generation number as measure of time, and the average synthesis of the language on the y-axis. Shaded regions are the LOESS estimates supplied by the ggplot2 package in R.	154
3.13	Results of Simulation 7, 125-Agent Networks: On the x-axis is the generation number as measure of time, and the average synthesis of the language on the y-axis. Shaded regions are the LOESS estimates supplied by the ggplot2 package in R.	154
3.14	Topology:Size Interaction: The mean levels of reanalysis from generations 100 to 200 are shown with 95% confidence intervals.	157
5.1	Grammar at Generation 1. (Top) Rewrite Dependency Graph (Bottom) ARS Rules	182
5.2	Grammar at Generation 42. (Top) Rewrite Dependency Graph (Bottom) ARS Rules	184
5.3	Grammar at Generation 69. (Top) Rewrite Dependency Graph (Bottom) ARS Rules	186

Contents

Acknowledgments	i
Summary	ii
1 Introduction and Literature Review	1
1.1 Thesis Overview	4
1.1.1 Chapter 2 Simulations: Traditional Simulations and the High-Level Model	8
1.1.2 Chapter 3 Simulations: Synthesis and the Low-Level Model	9
1.2 Review of Central Topics	10
1.2.1 Social Structure, Intimacy, and the Clustering Coeffi- cient	11
1.2.2 Morphological Typology and Synthesis	15
1.2.3 Grammaticalization	18
1.2.4 From Free Morpheme to Bound Morpheme	24
1.2.5 Interdisciplinary Perspectives	29
1.3 Summary and Desiderata	30
2 High-Level Model	33
2.1 Traditional Models of Language Change	38
2.1.1 Simulation 1: Diffusion	40
2.1.2 Simulation 2: Diffusion Mechanics and Transmission	48
2.2 High-Level Model of Language Change	53
2.2.1 Mechanisms of Language Change	54
2.2.2 Language Model	56
2.2.3 Agents	59
2.2.4 Agent-Agent Communication	60
2.2.5 Intergenerational Transfer	64
2.3 Simulation 3: Network transitivity and linguistic reanalysis	68
2.3.1 Network Architecture and Communication	69

2.3.2	Prediction	71
2.3.3	Results	72
2.3.4	Discussion	74
2.4	Simulation 4: Network Topology and Linguistic Reanalysis .	77
2.4.1	Network Architecture and Communication	78
2.4.2	Prediction	81
2.4.3	Results	83
2.4.4	Discussion	83
2.4.5	Exploring Parameters	87
2.5	Simulation 5: Revisiting Network Size	90
2.5.1	Results and Discussion	91
2.6	Conclusion	95
3	Low-Level Implementation	99
3.1	Meaning and Signal Spaces	102
3.2	Grammar & Production	104
3.3	Language Acquisition	107
3.3.1	Segmentation	110
3.3.2	Modified GRIDS: Extension of Idea	116
3.3.3	Modified GRIDS: Abstraction	118
3.3.4	Modified GRIDS: Generalize	120
3.4	Processes of Morphological Development	123
3.4.1	Extension	123
3.4.2	Repetition Effects	124
3.4.3	Reanalysis	125
3.5	Measuring Synthesis	127
3.5.1	English & Lushootseed Examples	130
3.5.2	Worked Analysis	138
3.6	Simulations	145
3.6.1	Simulation 6: Low-Level Transitivity	150
3.6.2	Simulation 7: Low-Level Social Structure	153
3.7	Discussion & Conclusion	156
4	Conclusion	161
5	Appendix: Diachronic Low-Level Example	179
6	Appendix: R Reports and Model Details	187
7	Appendix: Pseudo-Code	200

Chapter 1

Introduction and Literature Review

Natural languages vary widely in grammatical structure, especially in the degree to which they make use of nested composition. A graduated divide in this typological space is the extent to which words themselves may exhibit compositional patterning. That is, languages vary in the degree on average to which words possess internal structural hierarchies (**morphological** composition), beyond their arrangement in the phrasal hierarchy of a sentence (**syntactic** composition). In some languages, such as Chinese, word-internal morphological structure is virtually nonexistent, whereas others, like Lushootseed (an indigenous, Salishan language of North America), have so much word-internal morphological structure that entire utterances can be a single, highly complex word.

For example, to express the concept of an object to sit on, English speakers can use words like chair, stool, seat, etc. Each of these words cannot be

analyzed into further subdivisions that carry independent meaning, that is, each word is composed of a single morpheme. In Lushootseed, though, a single word is used to capture all of these: səx^wg^wədil, which can be further analyzed into s-əx^w-g^wəd-i-l (Bates et al., 1994). The word is built up conceptually from the parts, s (physical object), əx^w (can produce), g^wəd (be crouched), i (change state without external force), l (change state gradually). That is, all together the word means something like object that allows one to assume a crouched position. While not as analytic as Chinese, English generally has few morphemes per word, whereas in a highly synthetic (sometimes called **polysynthetic**) language like Lushootseed a word with a single morpheme is exceedingly rare.

This idea of degree of morphological composition was identified early on by linguists looking for potential features with which to classify and compare natural languages (Gabelentz, 1894), and such attempts to classify and compare languages has become a subfield within linguists known as **typology** (see Comrie (1989) for an excellent introduction). In more modern times, the degree of word-internal composition has been defined quantitatively, termed the degree of **synthesis** of a language, and remains a foundational feature in typological description (Greenberg, 1960). While presented in more detail in the following section on synthesis, the informal idea is that if one has a large, representative corpus of a language, the degree to which a language exploits word-internal composition can be measured as follows: First, segment the corpus into individual words and count them. Then, if possible, segment those words into the morphemes that compose them, and count those. In an analytic language like English, many words would be composed

of a single morpheme, e.g. chair, while a synthetic language like Lushootseed would have many morphemes per word, as seen with the word for chair above. With these two counts - words and morphemes - one can divide the number of morphemes by the number of words. This average measure of morphemes per word is synthesis.

Beyond the use of synthesis as a typological tool, linguists have long noted that languages with higher measures of synthesis tend to be spoken by small groups, while languages with lower measures of synthesis tend to be spoken by large ones. For example, Lushootseed at its peak before the colonization of the Americas likely had no more than a few ten thousands of speakers (Cook & Heizer, 1968), while languages like English and Chinese have had much larger speaker populations for most of their histories, and today have many hundreds of millions of native speakers. Evans & Levinson (2009) goes on to explain that not only does synthesis seem connected to population size, languages with low synthesis measures have similar word-classes and grammatical structures while languages with high synthesis measures display a rich diversity in grammatical organization and semantic categories. Indeed, empirical evidence suggests the typological patterning that languages display may be connected to aspects of the social network of the speakers. A recent survey of the World Atlas of Language Structures (Dryer & Haspelmath, 2013) found that after controlling for family similarity and areal influence, population spread was highly correlated with the number of grammatical features marked by morphological means (Lupyan & Dale, 2010). Specifically, languages with smaller and more isolated speaker populations were found to make much greater use of morphology than those

with larger and more wide-spread populations.

As summarized in Evans & Levinson (2009), the effect of social demographics on language structure affects numerous areas, including phonology, morphology, syntax, and semantics. This thesis contributes to this emerging area of research by specifically examining how a typological feature, synthesis, is affected by social structure. In terms of demographic variables, it investigates the relationship between small and large population sizes, and their interaction with the clustering coefficient, which as explained below in detail, may be a proxy for the general intimacy of a social group.

In terms of modeling, it directly incorporates three key processes identified in the linguistics literature on morphological development and operationalizes them at both a high-level and low-level of structural detail. In addition, it also considers the complementary directions of language spread, diffusion and transmission, within the same language model.

1.1 Thesis Overview

In the thesis, two main models are presented: the high-level model, and the low-level model. An alternative nomenclature for these might be the abstract model and the detailed model, respectively, but as both are quite abstract compared to the detail required to describe natural languages, the high vs. low naming was kept to not accidentally mislead or surprise the reader. The distinction between high vs. low can be understood in respect to how they approach the notion of synthesis. The high-level model does not measure it directly, but rather measures the necessary conditions for higher levels of

synthesis to develop. As discussed in more detail in this chapter, at least three cognitive processes are required to occur repeatedly for the level of synthesis to increase. The high-level model represents these processes using probabilities, and tracks how many times on average they are able to occur on the same linguistic forms through generations of agents. As such, social network structures in which these processes can occur repeatedly present the possibility for higher levels of synthesis to develop. However, there is not guarantee that supporting these processes actually leads to increased word-internal morphological composition. Metaphorically speaking, the high-level model provides a birds-eye view of how social network structure **may potentially** interact with language structure.

In contrast, the low-level model is run using the same network topologies as the high-level model, but is designed to measure synthesis itself. It does so by increasing the level of detail of the linguistic forms used by the agents. In the low-level model, the agents use linguistic forms with compositional structure, such that words and morphemes can be distinguished, and synthesis measures made. In addition, the processes required for synthesis to develop are modeled directly on this more detailed linguistic structure, rather than depending on probabilities. Again, metaphorically speaking, the low-level model provides a worms-eye view (so to speak) into how the social network structure **actually** interacts with language structure.

Traditional Models of Language Change

Simulation 1

Examines clustering coefficient in random network

Diffusion ONLY

Simulation 2

Examines clustering coefficient in random network

Diffusion AND transmission, but

Transmission change **same as diffusion** change

Figure 1.1: Summary of Simulation 1 and Simulation 2.

High-Level Model of Language Change

Simulation 3

Examines effect clustering coefficient in random network

Simulation 4

Examines effect of hierarchical vs complete topologies

Simulation 5

Examines effect of increasing magnitude of population

Figure 1.2: Summary of Simulation 3, 4, and 5.

1.1.1 Chapter 2 Simulations: Traditional Simulations and the High-Level Model

Chapter 2 presents 5 simulations in order to explore the ability of different social network topologies to support the processes required for higher levels of synthesis to emerge. Again, these simulations do not measure synthesis itself, but rather the potential capacity for synthesis to develop. The first two simulations, summarized in Figure 1.1, use a simplified model of language as often seen in the literature on language innovation. The purpose of these first two simulations is to illustrate the unexpected results when the distinction in learning mechanism used to spread innovations within a single generation and between multiple generations is ignored. The next three simulations, summarized in Figure 1.2, develop a novel approach (the high-level model) to remedy the problems motivated by Simulation 1 and Simulation 2.

More specifically, Simulation 1 presents results qualitatively similar to those in the literature on horizontal language change. Simulation 2 then adds a vertical dimension to the model, but does not account for the difference in horizontal vs. vertical language change. That is, in both simulations, language change is modeled simply as an all-or-nothing process in which an innovation is either acquired or not; there is no attempt to capture the gradual changes characteristic of vertical language transmission. In addition, both simulations are run on a number of networks with different clustering coefficients. This allows for the effect of this variable (a focus of this thesis) on such conventional models where it is typically ignored to be made evident.

Simulation 3 revisits the same network as Simulation 2, but with the new,

high-level model that captures the distinction between horizontal and vertical change. Moreover, it decomposes vertical change into the three necessary process for synthesis to develop. However, the network structures used in the simulations so far, while varying the clustering coefficient, ignore some other characteristic differences between small, intimate societies and larger, less intimate ones. Simulation 4 addresses this issue by using two network structures that more veridically capture this contrast. Finally Simulation 5 scales the networks from Simulation 3 and Simulation 4, each consisting of 25 agents, by an order of magnitude to 125 agents each, in order to examine the effect of size on the topologies investigated thus far.

1.1.2 Chapter 3 Simulations: Synthesis and the Low-Level Model

Chapter 3 presents 2 sets of simulations to investigate if an increased capacity for synthesis actually results in higher levels of synthesis itself. Simulation 6 used the same random network topologies as seen in Simulation 3 of the high-level model, in which the effect of the clustering coefficient alone was observed. In addition, Simulation 6 includes both 25-agent and 125-agent versions of the networks to be compared side-by-side. Then, Simulation 7 revisits the two network topologies that more veridically capture human social structure from Simulation 4 of the high-level model. As with simulation 6, both 25-agent and 125-agent versions of these topologies are included.

1.2 Review of Central Topics

As suggested briefly in the introduction, there appears to be a close connection between the structure of a language and the social demographics of its speakers. Amongst the numerous social and structural variables that could be investigated, this thesis will single out social “intimacy” as potential key factor in language change. However, as will be examined shortly, achieving a quantitative definition of intimacy is challenging. First, to help remedy this difficulty, a connection between the intimacy of a society and its **topology**, that is, the pattern of connectivity that characterizes the relationships between speakers, will be presented as a quantitative proxy. Specifically, a measure known as the clustering coefficient will be presented.

Next, the evolution of morphological typology as a subfield of linguistics will be presented, and the measure of synthesis will be formally given. As synthesis measures the average internal composition of words, the general linguistic account of how such composition emerges in natural language will be reviewed. As we will see, a key aspect of compositional development is the interplay between **diffusion**, the spread of language innovations horizontally across a network, and **transmission** the internal, structural innovations that occur vertically through the generations of language acquisition (Labov, 2007; Campbell, 2013).

Once this literature review is complete, the key findings will be summarized. From this summary, a number of desiderata for a model will be proposed, and used to motivate the design of the high-level and low-level models of the subsequent two chapters.

1.2.1 Social Structure, Intimacy, and the Clustering Coefficient

The role of network structure in linguistics has typically been dealt with in the subfield of sociolinguistics. In particular, the focus has been on social networks as a borrowed concept from sociology. The social structure of the agents is modeled as a mathematical object called a **graph**, and not to be confused with the more conventional English usage of the word, i.e. as a visual representation of data. The mathematical graph is a formal way of representing things and the connections between them - visualize a sheet of paper with black dots, representing the things, and lines drawn between them, representing the connections. Formally, a graph is specified by two components, called its **nodes** and **edges**. The nodes is a list of the things modeled by the graph (the black dots from the example), and the edges is a list of the connections between the nodes (the lines from the example). Formally, the edges are specified as pairs of nodes, where a connection exists between them.

To clarify further, let us assume we have a formal specification of a graph, i.e. we have the nodes and edges, and want to create a visual representation as described above. A straightforward procedure is to take all the nodes, assign a number or label that identifies each, and draw these out evenly spaced in a circle. Then, go through each pair of nodes in the edges, and draw a line between the two.

In the Sociological framework the structural aspects, at least in terms of complex network theory, have been downplayed in respect to the sociological

variables, such as power, race, economic status, etc. that delineate members of a community. However, Milroy & Milroy (1985) introduces the social network as a mathematical object for linguistics study, with speakers as nodes in a graph, and links structural variables to language variation. The paper builds deeply on Granovetter (1977), which is critical of the importance played in sociological network theory by strong interpersonal ties, where a so-called **strong tie** is defined as, “a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and reciprocal services which characterize the tie.” Conversely, a **weak tie** is one that scores low in these categories. The argument in both papers is that when examining the actual ties that connect individuals, i.e. producing a physical graph of the individuals and connections, the strong ties are readily apparent; however, the weaker, more peripheral ties that bridge gaps between groups of strongly connected individuals are crucial for exchange of information. Continuing to build on their examination of ties, Milroy (1987) presents **density**, defined as the ratio of extant ties in a network to the total possible ties, as a diagnostic variable, and Milroy & Milroy (1992) connects low-density networks with ease of spreading innovation and high-density networks with resistance to change (Chambers & Schilling, 2013). The reason for this is that in a high-density network, each person has many ties to other agents, and so unless an innovation enters the network from multiple directions at once, there is a far greater number of speakers using the non-innovative form. That is, in dense networks, the sense of normal vs. novel is greatly reinforced by the amount of related input any given person receives. Conversely, in a less dense network, the this sense of normal vs.

novel is not as strong, and the likelihood a novel lexical item may be adopted is increased.

Beginning in Trudgill (2002) and significantly expanded on in Wodak et al. (2010) and Trudgill (2011), five sociolinguistic factors that connect with language change are identified: population size, connection density, social stability, degree of shared background information, and degree of isolation. Trudgill (2011) presents the idea that ‘intimacy’ may be an important factor for developing complexity, and that it may be identified with dense, strong ties in networks. While earlier work suggested that strong ties are responsible for the inhibition of innovations, which is not disputed, it is proposed that these intimate connections provide the conditions necessary for creating language complexity, in particular bound morphology. On the one hand, variation must be present to create the conditions for reanalysis, but on the other, forms must be sufficiently frequent as to warrant the change. In networks of intimates there is the freedom and shared knowledge to create the variation, as well as the amount of contact to cement changes; thus changes, while more rare in dense social networks, will not easily be replaced once they do occur, and are more likely to continue to accrue additional complexity.

Finally, the notion of density in the social network is mathematically connected to the notion of ‘intimacy’ as proposed in Trudgill (2011). Beginning with Rapoport (1977), a probability argument was posed suggesting that if there are three individuals (A, B, and C) and further if strong ties connect A to B, and A to C, then almost certainly a weak tie exists or will soon form between B and C. That is, given three individuals and two strong

connections, a third connection that completes the triangle of connectivity will form. In a small, intimate society, there are many such strong connections (for instance, owing to family) and so one would expect the network structure to be composed mainly of triangular connections.

The clustering coefficient, also referred to as the transitivity of a network, can be measured in two ways: globally, reducing the entire network to a single density measure, or locally, capturing the density surrounding an individual. The global measure is defined as the number of closed triplets divided by the total number of triplets. A triplet is a unique set of three individuals and two connections that link them together in the network. For example, given the individuals (A, B, and C), if A is connected to B and B is connected to C, they would constitute a triplet. A closed triplet is a unique set of three individuals, and three connections that link them together, i.e. a triangular relationship amongst the agents. Returning to the previous example with A, B and C, they would form a closed triplet if additionally there was a connection between A and C. The local measure, defined for an individual agent, is the number of connections that **actually** exist between an individual's neighbors, divided by the total number of possible connections that **could** exist between the neighbors.

To summarize, the global clustering coefficient = $\frac{\#closed\ triplets}{\#all\ triplets}$, which characterizes an entire network, whereas local clustering coefficient = $\frac{\#existing\ edges}{\#possible\ edges}$ for a particular individual in the network. It should be noted that in this thesis the terms transitivity and clustering coefficient will be used interchangeably, but in particular the term transitivity will be favored in figures to conserve space.

1.2.2 Morphological Typology and Synthesis

Morphology, as a means of classification, was present at the beginning of modern typological study. In fact, before Gabelentz (1894), the field of inquiry was conventionally referred to as ‘the morphological classification of languages’ (Bakker, 2010). The Schlegel brothers first made the distinction between languages that utilize composition within words (von Schlegel, 1808) and those that lack such composition (von Schlegel, 1818). Shortly thereafter, Von Humboldt (1823) reported that languages exist in which entire utterances can consist of a single word form, and an attempt was made in Schleicher (1859) to arrange these into a classification scheme based on degree of morphological composition (Croft, 2002).

The modern vocabulary used to describe morphological types comes from Sapir (1921) in which he summarizes morphological type along two dimensions, namely, the quantity of morphemes and the manner of their combination. In terms of quantity, he distinguishes three classes: **analytic**, **synthetic**, and **polysynthetic**. Analytic languages are characterized by, on average, having about one morpheme per word, synthetic languages by having a few morphemes per word, and polysynthetic languages as having many morphemes per word. There is no clear limit to how many morphemes are required to distinguish synthetic vs. polysynthetic languages, nor does an analytic language always have one morpheme per word; the concept simply captures a general, graduated trend. As an additional criteria to help distinguish between synthetic and polysynthetic, he suggests that polysynthetic languages often have words with more than one ‘root’ morpheme,

where a root conveys basic semantic meanings and non-roots add more abstract grammatical information such as tense and agreement. In terms of manner, he distinguishes four classes: **isolating**, **agglutinative**, **fusional**, and **symbolic**. Isolating languages combine morphemes without affixation (i.e. with syntactic means only), agglutinative languages affix morphemes transparently, fusional languages affix morphemes non-transparently, and a symbolic language uses suppletion during affixation. Again, the notion of transparency is not precisely defined, but generally refers to the ability to which the individual morphemes can be distinguished in a word form. To clarify, Japanese or Turkish are examples of agglutinative languages, in which the shape of a morpheme changes little during affixation and is easily recognized in word forms. On the other hand, fusional languages are like Latin, in which a morpheme may undergo such phonetic change during affixation that it is not readily recognizable, sometimes requiring grammatical context to resolve ambiguity. An example of a symbolic language would be Ainu, in which, for a minority of verbs, a plural subject is signaled by a suffix, but the majority use an entirely unrelated word form for the plural (Refsing, 1986).

Sapir (1921) also makes it clear that for some languages a particular classification in terms of quantity or manner generally applies to the language, but for many others these terms apply only to a subset of the grammar and/or there may be a mix of types. For example, Japanese verbs are generally synthetic while nouns are isolating, or in Latin where derivational morphemes affix in an agglutinative manner while inflectional morphemes are fusional. While not having a particular solution to this issue, Sapir

makes this limitation of typological classification plain.

While perhaps known best for his later work on language universals and word order, Greenberg (1960) provides a solution to Sapir’s problem of applicability. He does this by developing 10 ‘typological indices’ that can be quantitatively measured given a corpus. The first two, the index of synthesis and index of agglutination, are specifically developed to capture the quantity and manner classes of Sapir. The index of synthesis was designed specifically to measure the distinction between analytic-synthetic-polysynthetic spectrum, while the agglutination index was designed to measure the isolating-agglutinative-fusional spectrum. Of interest for this thesis is the **synthesis** measure, which is formally defined as number of morphemes divided by the number of words in a corpus. In other words, $synthesis = \frac{\#morphemes}{\#words}$ for a given segmented corpus.

To provide a sense for this measure, some natural languages for which large corpora exist and the synthesis has been measured is presented in Table 1.2.3. At the top of the table is West Greenlandic, a polysynthetic language of the Eskimo-Aleut family, and at the bottom is Vietnamese, an analytic language of the Austroasiatic family. Vietnamese, similar in typology to Chinese, has a nearly minimum value of synthesis; a language in which all words are composed of a single morpheme, the synthesis measure would be exactly one. By way of comparison Modern English, while also analytic, does possess some bound morphology. For instance, verbs take the -s suffix in the third person present, regular verbs take -ed in the past tense, and -ing in the present progressive. Conversely, there is no absolute maximum value for synthesis, but the value for West Greenlandic is representative (Bauer,

2003). This may seem counterintuitive, for example, as one-word utterances with 10 morphemes occur naturally in West Greenlandic, and there is not strict limit to how many suffixes a verb may take Fortescue (1984). However, in contrast to these highly synthetic examples, many verbs only have a 2-3 morphemes, and the other constituents of multi-word utterances (e.g. nouns and articles) have 1-2 morphemes. Thus in a large corpus, the average measure of synthesis is much lower than may be intuitively expected.

1.2.3 Grammaticalization

Within linguistics, the creation of bound morphology is generally studied in connection with **grammaticalization**, the process by which grammatical forms emerge from lexical forms ¹. Like many central constructs within linguistics (cf. word), there is no standard definition of what distinguishes grammatical and lexical forms, although there is general consensus about intuitive, prototypical examples of each category. In general, ‘lexical’ forms serve to reference objects, actions, and qualities, whereas ‘grammatical’ forms serve to coordinate the mechanics of a language, such as agreement marking (Hopper & Traugott, 2003). For example, the English word “chair” serves the purpose of expressing the concept of an object to sit upon, and is prototypical lexical form. In contrast, the word “the” is a prototypical grammatical form, as it serves the abstract purpose of expressing the definiteness of the noun it agrees with.

¹In the following, the word form will be used to describe notions that broadly apply across levels of linguistic organization. That is, phenomena that apply equally to whole collocations, constructions, words, morphemes, etc. will be labeled ‘forms’, while specific terminology will be used when a phenomenon is more restrictive in scope.

Language	Synthesis
West Greenlandic	3.72
Sanskrit	2.59
Swahili	2.55
Old English	2.12
Lezgian	1.93
German	1.92
Modern English	1.68
Vietnamese	1.05

Table 1.1: Synthesis values for a small selection of natural languages. The minimum value for synthesis is 1.0 - a language in which all words are composed of a single morpheme. Conversely, there is no such mathematically defined maximum. (Adapted from Haspelmath & Sims (2013))

In principle, there is no reason that grammatical meanings need to be expressed with bound forms, nor that lexical meanings with unbound forms. An extreme case is the so-called lexical suffixes, an areal feature of many indigenous language of the Pacific Northwest, are added to roots purely to suggest semantic content. For example, in Lushootseed the suffix *-ucid-* can be added to a verb to express the notion of using the mouth or lips; it serves no grammatical function (Bates et al., 1994). However, many languages do tend to express ‘core’ grammatical meanings such as tense, number, case, etc. with non-free elements, such as bound morphemes or clitics, and in addition, it is common cross-linguistically that comparable grammatical meanings appear to develop from similar lexical sources; a compilation of several hundred are collected in Heine & Kuteva (2002). For instance, many languages that have a grammatical concept of future tense mark it with morphology that historically developed from a lexical form expressing volition or linear motion (Revere & William, 1994), cf. the English auxiliaries *will* and the more colloquial *gonna*.

The recurrence of such semantic/structural pathways in languages that are genetically unrelated (in the historical linguistics sense), resulted in the discovery and description of these so-called **grammaticalization clines** to become a focus in the literature, a nice summary of which can be found in Luraghi & Bubeník (2010). Entire volumes have been dedicated to such cross-linguistic clines, including indefinite pronouns (Haspelmath, 2001), spatial morphology (Svorou, 1993), temporal adverbs (Haspelmath, 1997), possessive markers (Heine, 1997), auxiliaries (Heine, 1993), and the morphology of tense, aspect and modality (Revere & William, 1994). Attempts to generalize the commonality of such clines has led to **unidirectional hypothesis**. While not having an accepted formal definition, it generally asserts the changes seen in individual clines are specific cases of a more general pattern of transformation: content item $\mapsto$ grammatical word $\mapsto$ clitic $\mapsto$ affix (Hopper & Traugott, 2003), and similarly, discourse $\mapsto$ syntax $\mapsto$ morphology $\mapsto$ morphophonemics $\mapsto$ zero (Givón, 1979).

As a general linguistic trend, the unidirectional hypothesis is able to account for a wide range of phenomena extending far beyond those mentioned just above. However, as is often the case with any so-called language universal, upon further investigation what actually exists is a strong trend rather than an rule without exception. Indeed, counter examples that seem to move in the opposite direction of the unidirectional hypothesis have been found, and the concept itself dubbed **degrammaticalization** to highlight the seemingly reverse directionality (Ramat & Hopper, 1998).

Norde (2009) presents a summary of the debate, in which the most widely cited example is that of the English genitive suffix -s (orthographically -'s);

in the sentence "the dog's bone" the -s on dog is used to show possession. While in Old English this form is clearly a bound suffix of nouns, it has expanded in scope. For example in Modern English, it is permissible to say "The dog that lives in the old house's bone is in my yard" where the possessive -s is semantically connected to the dog (the bone belongs to the dog, not the house) but is separated from the noun. This behavior is in an intermediate state between affix and grammatical word; it clearly is not bound the noun dog, but depends on it for meaning. Thus, we have a case where a grammatical maker, at one time undeniably a nominal suffix, has moved in the opposite direction of the unidirectional hypothesis.

Traditionally many such intermediate items have been identified as "clitics", but term is not yet defined in a universally accepted way. Klavans (2018) presents an excellent historical overview of the issues. The crux of the debate hinges on the notion of phonological dependence of the clitic to its host. Klavans (2018) argues that phonological dependence is simply one of many features such intermediate items progress through on the journey from free word to affix, and proposes that any such intermediate item be termed a clitic. This seems to be generally accepted in the literature on grammaticalization, although the notion of phonological dependence is the norm outside of this. In this thesis, the term clitic will be used in the more permissive sense of linguistic form occupying an intermediate position between free word and bound affix.

In response to such examples of degrammaticalization, Heine (2003) provides an overview of the main defenses. First and foremost is the reassertion that the unidirectional hypothesis, like all language universals, are not to

be taken as physical laws. Instead they offer indications of strong general trends that help linguists understand cross-linguistic commonalities, and potentially something deeper about human language itself. Second, they argue that while indeed there are cases, like the English possessive marker, in which forms have shifted contrary to the unidirectional hypothesis, no forms have ever fully progressed on the cline in the contrary direction. That is, they argue that there are no cases of affixes gradually progressing into clitics, then into free grammatical words, and finally into free lexical words. Again, all the examples of degrammaticalization involve forms the began as predicted, as free words, despite the evidence of some retrograde tendencies.

The ongoing debate about grammaticalization has led some linguists to posit that perhaps the reason such degrammaticalization counter examples exist, is that more general phenomena of language change are interacting at multiple levels. That is, what exists is not a single grammaticalization phenomenon, but rather the observation that cross-linguistically patterns of change that follow the unidirectional hypothesis exist because of a subset of domain general, cognitive processes Newmeyer (2000). However, this argument can be taken to the extreme, such as in Campbell (2000) where it is suggested that idea of grammaticalization and the unidirectional hypothesis be dispensed with entirely, for if it is a composite of domain general processes, then it really has no ontological status. This would seem, as the phrase goes, to throw the baby out with the bathwater. Rather, the position taken in this thesis aligns with that expressed in Fischer et al. (2004), where it is acknowledged that while grammaticalization (and by implication, degrammaticalization) are not holistic entities, the fact that such patterns

occur cross-linguistically and in a general progression is valid area of research. Further, the fact that grammaticalization can be better analyzed as a set of interacting subprocesses only stands to strengthen the field as such finer details are investigated.

The holistic concept of grammaticalization, I would like to point out, minimally conflates a number of distinct linguistic dimensions. Moving along the unidirectional hypothesis stages from left to right, the forms increase in boundedness, increase in grammaticality, increase in frequency, decrease in semantic specification, and become increasingly obligatory. As suggested in Fischer et al. (2004) and more modern treatments of the subject, such as Narrog & Heine (2011), such dimensions are best dealt with individually, as orthogonal processes, before being considered in unison. This strategy is precisely what is pursued in this thesis, singling out the dimension of boundedness for investigation.

Fortunately for the discussion of bound morphology, the mechanistic aspects of grammaticalization have already been discussed in terms of three specific sub-processes. That is, regardless of ontological status, i.e. whether they form a holistic framework called grammaticalization or are a collection of domain-general process, the individual stages can be discussed free from such concern. Indeed, there is no attempt in this thesis (or in the design of the models) to represent concepts such as lexical vs. grammatical meaning, or derivational vs. inflectional morphology. Instead, the focus will rest squarely on the process identified in the creation of bound morphemes, and developing models that help assess their interaction with social structure. In the next section these three process will be identified,

and the manner in which they interact to produce bound morphemes from free words will be discussed.

1.2.4 From Free Morpheme to Bound Morpheme

Within grammaticalization, three general mechanisms of change are traditionally distinguished: **extension**², **reanalysis**, **repetition** (Narrog & Heine, 2011). Extension refers in general to any use of language that results in a linguistic form being used in a novel, unconventional manner. For example, semantic extension is often caused through the use of analogy or metaphor; for example, the word *autumnal* literally means having to do with Autumn, but by extension describes things in the later stages of their lives. Similarly, morphological and syntactic extension is common through the use of existing patterns with novel constituents, such as the use of ‘email’ or ‘Google’ as verbs based on existing pattern of other nouns that serve as tools (c.f. *hammering*, *stapled*, *paper-clipped*).

Reanalysis refers to changes in the structure of a linguistic form from one generation to the next, whether in terms of meaning, phonology, or grammar. For example, semantic reanalysis results when the meaning of a word shifts: the use of ‘wicked’ at the dawn of the 20th century strictly meant things of an evil nature, but today it can also be used to refer to things that are good, or awesome. A nice example of structural reanalysis in English can be seen in the change from ‘a napron’ to ‘an apron’ where the form of the definite article was reinterpreted. The above examples focus

²Originally called *analogy* (Meillet, Meillet; Heine, 2003), the term has begun shifting in the literature to *extension* in order to accommodate more general changes of kind (Himmelmann, 2004; Narrog & Heine, 2011)

on semantic and overt structural changes, but extension and reanalysis are not restricted to just these domains.

The final mechanism, repetition, refers to changes caused by a form becoming routinized by speakers due to repeated use. In particular, **phonetic erosion** happens when a form is so predictable from context that speakers gradually reduce the phonetic realization of a form, eventually omitting phonological material; the contraction from *I shall* and *I will* to *I'll* are examples. In addition, **chunking** is a more covert effect of repetition in which forms that co-occur so frequently get routinized into an essentially inseparable unit. It should be noted that often such contractions have the apparent effect of reducing the morphological complexity, and indeed with strict respect the form in question, they do. However, as in the English example, both the full form and reduced form often continue to exist as alternatives for generations. In addition, it is not the case that the use of “will” has been lost from English grammar, only that an additional form has been added, which as explained below, provides the potential for future reanalyses.

The traditional account of the unidirectional hypothesis relates these mechanisms together to explain the gradual, directed change from free, lexical forms to dependent, grammatical ones. As forms are extended, learners typically attribute a more general meaning to the signal than was originally understood by prior generations of speakers. For example, and one we will consider in more detail shortly, the Proto Indo-European (PIE) future developed from a desiderative; the notion of desire for an action was extended to imply a future time frame for the action. However, when such exten-

sions enter their input, new learners occasionally internalize the extended meaning as primary, eliding the original meaning. In the linguistics literature, this kind of semantic change that makes a signal more general through repeated reanalysis of generalized contexts is called *bleaching* (Hopper & Traugott, 2003), as in some sense the colorful patterns of the signal have been transformed into a generic, white mass.

When a signal is ‘bleached’ in this manner, its frequency of usage may increase, as its more general meaning can be applied in an increased range of communicative situations. Speakers can more precisely express themselves without implications introduced by extending alternative forms that are less bleached, and thus have additional connotations. This increase in usage makes the bleached form more predictable, as it occurs more and more frequently in similar contexts, and it becomes a target for repetition effects. As phonetically reduced forms are reanalyzed by subsequent generations as underlying, the form becomes more likely to be structurally reanalyzed as a dependent element of other signals, usually first syntactically, and then eventually morphologically.

As a schematic, we can see this transition as the interaction of those three general processes as follows. First, extension works in tandem with semantic reanalysis (bleaching) to make a form increasingly general and applicable over several generations. This has the potential to boost the frequency of the form, making it the target of repetition effects. In particular, it becomes reduced in phonetic form and chunked with its neighbors. Again reanalysis works in tandem, causing the form to become reanalyzed as a clitic if the neighboring element it has been chunked together with is a phrase, or

reanalyzed as an affix if the neighboring element is a word. In many cases, the continued reduction in phonetic expression causes a cliticized element to further bind to an individual word class as an affix.

A classic example of the entire transition is the history of the future tense from Proto-Indo-European (PIE) down to the Romance languages, here exemplified by Spanish. As discussed in detail in Giannakis (1993), in PIE there was no future tense; however, from the PIE desiderative, Classical Latin developed a dedicated future inflection, e.g. Latin *amabō* ‘I will love.’ However, this inflection was lost by the time of Vulgar Latin, and the use of the auxiliary verb *habere* ‘to have’ became conventionalized to express the future tense; ‘I will love’ had become *amare habeō*, or ‘I have to love.’ In the individual Romance languages, this convention followed language-specific phonological changes. The progression in the ancestor of Spanish resulted in **amar he* and eventually the auxiliary verb became the future inflectional endings of modern Spanish, *amaré*. Here the process began with an existing PIE inflectional suffix being extended for a new purpose in Classical Latin, but over time, repetition led to phonetic reduction, and eventually it was reanalyzed as not being present at all. Then a typical lexical item, the verb ‘to have’, is extended to fill this void. It too is progressively reduced in phonetic form – *habeō* $\mapsto$ *he* $\mapsto$ *-é* – while losing semantic content (bleaching) and becoming more restricted in compositionality; it progressed from a free, general-purpose verb to a bound verbal inflection, simply expressing the future tense. It is likely that this inflection too will be lost at some point in time, and the process will begin anew.

This decomposition of the transition from free morpheme to bound, sheds

light on its close connection to grammaticalization. Typical grammatical forms express fundamental notions that cross-cut the majority of communication. For example, the time of an action is often a relevant piece of information. As such, it is natural for a means of expressing this concept to emerge in a language. However, once a form for expressing time becomes conventionalized (c.f. the use of ‘to have’ in the PIE example) it becomes a target for repetition effects, that in conjunction with reanalysis, lead to bound morphology. However, in the case of grammatical concepts like time, aspect, number, etc. the means of expressing them are already highly frequent due to their core meanings, and thus even more likely to become bound morphemes, and eventually so phonetically reduced that they are lost. There is nothing special about grammatical vs. lexical forms at work, simply the frequency of the concepts they tend to represent. In fact, divorcing the notion of grammaticality from these subprocesses allows one to explain a greater range of phenomena. For instance, the derivational morphology example of the Vulgar Latin prefix ‘pre-’ makes perfect sense in this light. The notion of coming before is a useful distinction, but not as central to all communicative events as tense and aspect. The same processes lead it to becoming a bound affix from an earlier Latin adverb, *prae*, without any need to appeal to the problematic grammatical vs. lexical distinction. However, it is not so frequent in communication that it is a target for continued phonetic reduction or further chunking, which would eventually lead to loss. It has remained a stable piece of derivational morphology down from Vulgar Latin into the individual Romance languages, while other inherited items of a more frequent and grammatical nature have undergone drastic changes or

loss.

1.2.5 Interdisciplinary Perspectives

Within linguistics, the study of bound morphology is closely tied to the distinction of grammatical from lexical items. However, the study of why human language should exploit combinatorial, compositional structure, of which syntax and morphology are specialized cases, has long been studied under the interdisciplinary umbrella of cognitive science. Smith et al. (2008) provides a current overview of major issues and approaches.³

In particular, the concept of human language as a product of **cultural evolution** has been a focal point (Smith et al., 2003). The distinction between cultural evolution and conventional, biological evolution is that each learner of a language must observe the linguistic behavior of others and infer the structure of the language from this input; one does not inherit in their DNA grammatical patterns or lexical knowledge. We do seem to inherit endogenous abilities, though, that interject a bias towards certain analyses (Han et al., 2016). As a consequence of this, small changes in the input of language learners can lead to novel patterns in one generation that were not found in the previous (Grünwald, 2007; Langley & Stromsten, 2000; Senghas

³There is some variation in the usages of these terms. Hockett & Hockett (1960) states that ‘duality of patterning’ is an essential design feature of human language. The term duality comes from the fact that two levels of ‘combinatorial’ structure of introduced, one meaning-less and the other meaning-full. The example given is phonology, a system in which components combine given established patterns but are essentially meaningless, as compared to morphology and syntax, in which the combination of elements lead to predictable meanings. In order to better distinguish between these two notions of structure from parts, De Boer et al. (2012) suggests **combinatorial** when referring the former, and **compositional** when referring to the later. This convention will be used in this thesis. In addition, sometimes the term **hierarchical** is used contrastively to refer to combination with meaning.

& Coppola, 2001). On this account, the observed structural changes in a language as seen over historical time are the result of small, incremental changes in the input to successive generation of learners.

One main thread of thought is that combination and composition emerge in response to communicative pressures inherent in cultural evolution. From the perspective of a diffusion, human languages cannot be too simple or too specific as to impede expression of unpredictable and novel meanings. From the perspective of transmission, human language must be capable of being mastered in a short period of time, as well as being learned from a small sample of usage examples. Composition and combination provide solutions to these problems. When meaning and signals have a combinatorial structure, their constituent components can be reused, which reduces the burden of learning and use (Zuidema & De Boer, 2009; Nowak et al., 1999). Similarly, the composition of linguistic signals allows for them to form boundaries and classes. These patterns make the processing of language easier, as well as further reducing the burden of learning by allowing the combinatorial units of language to be categorically organized (Smith & Kirby, 2008).⁴

1.3 Summary and Desiderata

Summarizing the above literature, a number of important considerations must be accounted for in a model of morphological complexity. Further, the literature suggests a connection between a mathematical property of social

⁴There is, naturally, a debate as to if combination and composition are not in fact the same process acting on different structures. While immaterial for this thesis, the reader is directed to Smith et al. (2008) for a diversity of viewpoints.

network structure, and the way in which language complexity forms.

Beginning with model desiderata, first, the distinction between transmission and diffusion must be present. Vertical transmission is generally responsible for the incremental increase of structural complexity. In contrast, diffusion places a different pressure on forms for ease of use, and can lead to competition in which more complex developments are replaced by structurally more simple ones (Reali et al., 2018).

Second, there are three well-studied mechanisms, namely, extension, re-analysis, and repetition, which lead to the formation of bound morphology. In particular, extension and repetition effects are driven by fully competent speakers of a language, and set the stage for potential structural changes to be actualized by language learners. It is the repeated recurrence of this pattern that eventually drives free forms to become bound forms. It should be noted that these processes do not depend on the grammatical/lexical distinction of grammaticalization theory under which they were generally developed.

Third, interdisciplinary work in cognitive science has suggested that compositional patterns (as seen in syntax and morphology) emerge in response to the communicative pressures of cultural evolution on language. It identifies change in the input and incomplete information as the key for introducing new compositional units. This complements the more intricate account from the linguistics literature, which essentially asserts the same basic notion: transmission of incomplete information leads to structural change.

These findings on morphological development intersect with the concept of the social network as modulating innovation and language change. The so-

ciolinguistic literature has identified connection density, a measurable property of the social network as responsible for the acceleration and hindrance of linguistic innovation. In particular, dense, intimate connections hinder introduction of forms via diffusion, and create the conditions for slow, incremental change due to transmission. In contrast, sparse connectivity seems to favor the diffusion of innovative forms, albeit at the expense of replacing more complex variants already in existence (Milroy & Milroy, 1985).

While density can be quickly measured, the clustering coefficient is a better fit. Mutual, triangular relationships, which the clustering coefficient is a ratio, appear naturally in connection with strong ties. In turn, such strong ties are likely to be increasingly present in societies of intimates. This natural fit makes the clustering coefficient a structural proxy for the concept of intimacy without modeling emotional and familial connections.

Finally, the formation of bound morphology is dependent on the input that language learners receive. The intimacy of connections, as measured by the clustering coefficient, provides a direct link to how that information is distributed within the network. It provides a common point of reference for the sociolinguistics literature on network structure and the interdisciplinary work on bound morphology to intersect: social structures that favor transmission should favor the development of bound morphology, and typologies that utilize it. In contrast, social structures that favor diffusion should favor non-morphological means of composition. The clustering coefficient presents a continuous measure between the two.

Chapter 2

High-Level Model

In this chapter, the high-level, agent-based model of diachronic language change is developed. The chapter is composed of 5 simulations, which conceptually comprise 2 main parts: the first (covering Simulations 1 and 2) concerns traditional modeling of language change and potential issues therein with respect to synthesis. The second, (covering Simulations 3, 4, and 5) proposes a novel model that better accounts for the features essential for addressing synthesis. All the simulations in this chapter do not measure synthesis directly; none of them are at a level of detail that distinguishes words and morphemes, which are necessary for the definition of synthesis. However, the high-level model used in Simulations 3, 4 and 5 measures the potential capacity for synthesis to develop.

As will be discussed in detail below, the high-level model measures this capacity by operationalizing the three processes identified in Chapter 1 as necessary for synthesis to develop: extension, reanalysis, and repetition. The processes are modeled as probabilistic occurrences, where the proba-

bilities are derived on an agent-to-agent basis from the network structure surrounding the agent and language experience of the agent during its ‘life span’ before vertical transmission. As discussed in Chapter 1, crucially important for the development of bound morphology is the repeated reanalysis of the same linguistic form where the effects of extension or repetition have increased its frequency or economized its expression. The high-level model tracks the number of times such co-occurrences of extension/repetition followed by reanalysis have been repeated through transmission; it does not track the number of time each individual process has occurred, only the instances where this particular combination of processes has been suggested in the literature to drive the creation of bound morphology.

The first part concerning traditional models (Simulations 1 and 2) is presented to motivate the creation of the high-level model. It does so by demonstrating qualitatively that failing to account for the differences between horizontal and vertical language change, as commonly done in the literature, leads to predictions that do not accurately predict extant observations of morphological development. In particular, Simulation 1 will capture commonalities found in several influential papers on language change, and demonstrate how they interact with the main variable of interest in this thesis, the clustering coefficient. In Simulation 1, only the horizontal aspect of language is present. Simulation 2 extends the traditional model of Simulation 1 by adding the vertical dimension of language change, but does not make any effort to account for difference in horizontal diffusion vs. vertical transmission; in both dimensions, only the presence or absence of change is recorded without any notion of gradualness, as would be and expected

Traditional Models of Language Change

Simulation 1

Model predicts high clustering promotes grammatical diffusion

Result is counterintuitive to natural language observations

Simulation 2

Model predictions are mixed

Result is counterintuitive to Simulation 1

Result is highly dependent on random aspects of connectivity

Figure 2.1: Summary of Results from Simulation 1 and Simulation 2.

aspect of vertical transmission.

In both Simulation 1 and Simulation 2, as common in the literature, linguistic constructions are modeled as tokens the agents can probabilistically pass to their neighbors in the network. As is also common, the distinction between morphologically complex and non-complex constructions is modeled by setting the probability a token is transmitted to be low or high, respectively. The idea being based on the cross-linguistic observation that constructions utilizing bound-morphemes are more difficult for adults to acquire through use than constructions with little or no bound morphology involved (Jiang, 2004). The results of these two simulations are summarized in Figure 2.1. Simulation 1 demonstrates that such modeling of grammatical

High-Level Model of Language Change

Simulation 3

Clustering coefficient positively correlated with support for synthesis
Minimum clustering value require to obtain highest levels of support

Simulation 4

Hierarchical topology supports synthesis at levels lower than predicted
Hub agents responsible for counteracting support

Simulation 5

Increases in size magnify effect of existing biases for support

Figure 2.2: Summary of Results from Simulation 3, 4, and 5.

forms as low probabilities leads to results that contradict extant observations. Similarly, Simulation 2 shows that the manipulation of the clustering coefficient yields unstable results incongruent with both Simulation 1 and natural language observations again. It is concluded then that such traditional models, while useful for demonstrating horizontal diffusion of lexical items, are not wholly adequate for capturing the dynamics of vertical transmission.

In the second part, covering Simulations 3, 4, 5, the high-level model is developed and presented in contrast to Simulations 1 and 2 prior. The results are summarized in 2.2. Simulation 3 uses the same network structure as Simulation 2, but finds results congruent with natural language obser-

vations, at least so far as the sustained support of the processes necessary for bound morphology development may predict synthesis. However, while the networks in Simulation 2 allow for manipulation of the clustering coefficient, suggested in Chapter 1 as a potential proxy for intimacy, they lack other features present in human social networks. Specifically, as social networks grow in size and become hierarchical, the distribution of clustering varies inversely with degree; that is, agents with many connections tend to exhibit little clustering around them, while agents with few connections exhibit high levels of clustering. Two networks are presented: the Complete network, used to model a small, intimate society, and the Hierarchical network, used to model a larger, less intimate one that possesses the right distribution of clustering within it. The main result is that the Hierarchical network possesses far lower levels of support for the processes identified as key for bound morphology, and further, the effect of hub agents are identified as the major contributing factor for these lower levels of support. Finally, while Simulation 4 examined the differences in connectivity between small and large societies as explanatory, it held the size of the populations constant at 25-agents, as was also done in Simulation 3. To better understand the effect of size, the networks from Simulation 3 and Simulation 4 are adapted to 125-agent versions. The main result is the apparent biases present in the 25-agent networks are amplified in their 125-agent counterparts.

2.1 Traditional Models of Language Change

As discussed in Chapter 1, there is an important distinction to be made between the diffusion of lexical innovations across a network of speakers, as opposed to the development of grammatical innovations through generations of language learners. In particular, new words and set phrases are borrowed by other speakers of the language through exposure, diffusing across the network as they are adopted. In contrast, transmission is the means by which new grammatical structures slowly emerge as the usage of speakers biases the input for reanalysis.

This is by no means a strict dichotomy. For instance, Seifart (2012) presents evidence of entire noun and verb agreement paradigms being borrowed in Arawakan languages of the Amazon, and similarly, Ralli (2012) presents examples of verb being borrowed from Turkish into a dialect of Greek while retaining their original Turkish inflectional pattern. However, this is not the common means by which morphology enters a language, and even more so, nor the way entire typological shifts happen. For example, in the case of extreme paradigmatic borrowing presented in Seifart (2012), the new paradigms were limited to the subset of borrowed words, and did not spread to other areas of the language. Similarly, the verbs in Ralli (2012) were borrowed with inflectional morphology, they were done so because the semantics of these inflections closely matched a tense distinction made in Greek verb stems. The inflectional morphology of the Turkish verbs, while transparent, was not analyzed as such; each of the inflected forms were borrowed as stem alternations, a pattern already existing in Greek.

As mentioned in both of these articles, while the morphology was transparently borrowed, there were already existing semantically or structurally similar forms in the borrowing language. We can see a similar pattern in English with respect to the large number of Romance (Latin, Old French) borrowings in the post Old English stages of development. Along with these vocabulary items came an entire system of prefixes and suffixes, however, they remained largely confined to the Romance vocabulary, and may be less transparent to native speakers. For example, consider the words “hydrate” and “hydration”, not at all restricted to academic literature in today’s health-focused economy. Each is arguably composed of minimally a stem *hydr-* and a suffix, *-ate* or *-ation*, where countless examples of such pairs exist in the English lexicon. However, the grammar of English treats these words holistically, as a verb and noun, respectively.

In any case, while it is clear that morphology **can** be diffused, what is of interest in this thesis is how that morphology was generated to begin with. In Seifart (2012) where the entire verb and noun paradigms were borrowed, or the derivational Romance morphology borrowed into English, the donor languages had to acquire that morphology from somewhere. What is at issue in this thesis and these models, is trying to understand how the social network structure affects the generation of morphology itself.

However, many extant models do not make a distinction between these two kinds of mechanisms (Baxter et al., 2006; Nettle, 1999; Ke et al., 2008; Ke & Yao, 2008), namely, adoption of novel lexical items through exposure (which may include morphology, as described above) and internal structural changes via repeated reanalysis. Moreover, while some current models do

incorporate a means of transmission, they often fail to use an additional measure more appropriate for measuring the incremental changes associated with transmission (Real et al., 2014, 2018). To clarify, in diffusion of forms, the form is generally borrowed or not borrowed. However, the changes that occur in transmission are incremental, and such that even if speakers of different generations are producing the same surface forms, the semantic understanding or structural interpretations may be subtly different. In order to track such incremental changes, a binary measure of existence/non-existence as useful for modeling diffusion should be augmented with a continual measure that records the repeated occurrence of these small changes in transmission.

To better illustrate why both diffusion and transmission must be measured appropriately, a comparable model to current studies will be developed in following two simulations. First a typical diffusion model will be created, wherein the agents will acquire innovations through repeated exposure. In the second simulation, a mechanism for transmission will be added, but the use of a diffusion based measure continued. In addition, the network structure used, while also commonly used, will allow for the variable of interest of this thesis, the clustering coefficient, to be varied. In this way, the relevance of exploring this parameter can be demonstrated.

2.1.1 Simulation 1: Diffusion

The effect of social network structure and language change has been studied computationally in several well-cited papers (Niyogi & Berwick, 1997; Baxter et al., 2006; Nettle, 1999; Ke et al., 2008; Ke & Yao, 2008; Real et al., 2014,

2018), and it is from these which the current model will generalize. That is, while each model in these papers is unique, they share a set of common design features, and this is what will be captured in the model used in these first two simulations.

Model

Common to all, is the modeling of language change as a diffusion process, whereby a particular form (or forms) spread from one agent in the network to another as a function of probability. Typically the probabilities are set such that repeated exposure is required before an agent will adopt the innovation.¹ In such models, the spread of the innovation as a function of time is measured, typically reporting the proportion of the agents having adopted the innovations.

As described in Chapter 1, the social structure of the agents is modeled as a mathematical object called a **graph**, and not to be confused with the more conventional English usage of the word, i.e. as a visual representation of data. The mathematical graph is a formal way of representing things and the connections between them - visualize a sheet of paper with black dots, representing the things, and lines drawn between them, representing the connections. Formally, a graph is specified by two components, called its **nodes** and **edges**. The nodes is a list of the things modeled by the graph

¹Most recently, Reali et al. (2018) use different probabilities, high vs. low, suggesting that grammatical innovations can be represented by low probability innovations while lexical innovations represented by high probability innovations. While certainly true that grammatical innovations such as bound morphology are less likely to spread through repeated exposure than lexical innovations, there are very few cases of grammatical borrowing in the linguistics literature, and even these required intergenerational transfer to be adopted (Heine, 2003).

(the black dots from the example), and the edges is a list of the connections between the nodes (the lines from the example). Formally, the edges are specified as pairs of nodes, where a connection exists between them.

To clarify further, let us assume we have a formal specification of a graph, i.e. we have the nodes and edges, and want to create a visual representation as described above. A straightforward procedure is to take all the nodes, assign a number or label that identifies each, and draw these out evenly spaced in a circle. Then, go through each pair of nodes in the edges, and draw a line between the two.

In terms of network structure, a variety of **topologies** (patterns of connectivity that characterize the structure) have been used these studies, but the most common is a **random network**. A random network is specified by two parameters, the **size** n and **connection probability** p (Erdős & Rényi, 1959). To construct a random network from these parameters, first a number of nodes are created equal to the size. Then, one-by-one, each possible pairing of two nodes is considered, and with a p chance an edge is created between them. The random network, in addition to having few parameters, has the property that the connection probability is approximately equal to the global clustering coefficient (Barabási, 2014), which is the variable of interest for this thesis and suggested proxy for intimacy.

The two simulations in this section will follow this established convention of modeling language change as a diffusion process. Each agent in the network will be in one of two states: having adopted the innovation and not yet having adopted it, and an agent having adopted the innovation will be termed an innovator. At each interaction between an agent that has

adopted the innovation and one that has not, there will be a set probability, the **diffusion probability**, that the innovation will diffuse; the agent will change its state to having adopted if it has not done so already. As is also common in diffusion models, there will be no probability that the an agent will give up the adopted innovation and return to the unadopted state.

To assess the effects of diffusion, 600 unique random networks were constructed and tested: for each of 10 connection probabilities (0.1 to 1.0 in +0.1 increments) 30 unique networks were constructed with 25 agents, and another full set with 125 agents; 10 probabilities * 2 sizes * 30 replications = 600 networks. Each network was allowed to run for 100 rounds of communication before being stopped. A round of communication consists of each connection in the network being considered once in random order, and allowing for the innovation to potentially diffuse if one agent in the pair has adopted the innovation and the other has not. The diffusion probability was set at $p = 0.01$. At the end of each round the total proportion of agents using the innovation was recorded. In order to begin, one agent is selected at random to become an innovator, while all others have yet to take on the innovation.

The size of the networks were set at 25 and 125 agents, due to a detail of construction of the Hierarchical network used in Simulation 4 and Simulation 5 of this chapter, and Simulation 7 of Chapter 3. Its structure is hand specified, i.e. it does not use an algorithm like the other networks in the thesis, and that construction entails using agent multiples in powers of 5 (Ravasz & Barabási, 2003). In terms of the diffusion probability, it is medial in magnitude to those used in the studies mentioned: 0.002 used in Reali

Figure 2.3: Simulation 1, 25-agent network results: the total average proportion of innovators is displayed on the y axis, and the communication round number as a measure of time on the x axis.

et al. (2014) is the lowest, and .75 in Ke et al. (2008) is the highest. However, in all of these diffusion models, given enough time, the entire network will eventually adopt the innovation; there is no means for an innovation to be lost. The diffusion probability then simply controls the quickness of innovation spread, all things else being equal.

Results and Discussion

The results of the simulation are presented in Figure 2.3 and Figure 2.4. At the end of each round of communication, the total proportion of innovators is measured and displayed on the y-axis. In both figures, each line represents the average measure across all 30 replications of the indicated topology.

Figure 2.4: Simulation 1, 125-agent network results: the total average proportion of innovators is displayed on the y axis, and the communication round number as a measure of time on the x axis.

As one might expect, the more connections there are in a network, the quicker the innovation will spread, as each connection is another potential change for diffusion to occur in the round. This can be seen in that both the 25 and 125-agent networks the number of rounds required for an innovation to spread decreases as the connection probability increases. Similarly intuitive, as the size of a network increases, the same connection probability yields more connections in the network, which increases the number of chances per round of communication that the innovation may spread. Taking a network with a connection probability of 1.0 as the most extreme illustrative example, the 25-agent network has 300 connections (edges) and the 125-agent version 7750 connections. Thus, while there are 5 times the agents in the 125-agent version that must adopt the innovation, the number of opportunities for the innovation to have spread between successive communication events is much greater.

Also, most visible in the 25-agent graph (Figure 2.3) is that equal increases in connection probability do not amount to equal gains in the number of rounds to complete diffusion. For example, the 0.2 network is completely diffused on average by round 75, the 0.3 network by round 37, and the 0.4 by round 25. This added effect of connection probability is due to the fact that connection probability is approximately equal to the clustering coefficient, and thus the number of triangular connections also increases.

This effect is easy to explain with an example. Consider three agents, call them A, B, and C, who are connected in a line; A is connected to B, and B is connected to C. Now, if A were to be an innovator, in order for that innovation to reach agent C, it must first successfully be passed to B. This

imposes two potential ways that the innovation can fail to reach C. First, A might fail to spread the innovation to B, and thus it will be impossible to reach C when it communicates to B. Second, in the communication round, B and C might communicate before B communicates with A, and thus even if B does adopt the innovation from A, it will have to wait until the next round to potentially diffuse to B. Next, imagine if the three agents are in a triangular connection, such that A is also connected to C. In this case, if A fails to spread the innovation to B, it still has a chance to spread directly to C. Alternatively, if A successfully spreads the innovation to B, there are now two chances for the innovation to reach C instead of one. As triangularly connected agents begin to share connections with other triangularly connected agents, this effect continues to compound.

Despite the intuitive behavior of the model, though, these results are counterintuitive to the expectations of the sociolinguistics literature. In the model we find that the networks with higher clustering seem to encourage the spread of difficult to adopt innovations, which some have tried to link with grammatical vs. lexical innovations, more so than those with less clustering. However, observations of language change in reality have demonstrated that dense connections are more resistant to change, in particular ones that involve morphology (Milroy & Milroy, 1992; Labov, 2007). For the creation of bound morphology, it is not important how quickly an innovation may diffuse, but how long that innovation may be retained and slowly altered by generations of speakers through transmission. That said, some models have added a transmission component (Nettle, 1999; Ke et al., 2008; Real et al., 2014, 2018), but continue to use diffusion processes as the measure of the

network to support innovation. In the next simulation, we will demonstrate why that is not an ideal solution for studying transmission related changes, i.e. morphological vs. lexical change.

2.1.2 Simulation 2: Diffusion Mechanics and Transmission

In order to observe the interaction effects of transmission on a diffusion model of change, the same networks from Simulation 1 will be rerun to incorporate a simple transmission mechanism. While the diffusion mechanics will be retained exactly as in Simulation 1, they will only be allowed to operate for a limited number of rounds, after which each agent will reset its state based on the state of those it is connected to. That is, the process of intergenerational transfer will be modeled based on the potential input the agent receives from those it communicates with. While still not yet modeling the processes specific to transmission (extension, reanalysis, and repetition mentioned in Chapter 1) we can see how the addition of even a simple, usage-based intergenerational transfer process affects the conclusions one might draw from the diffusion model above.

Model

As mentioned above, the rules of diffusion and probability to diffuse are retained exactly as in Simulation 1, but agents will change their state based on those they communicate with after 50 rounds of diffusion. This limit of 50 rounds was selected as all of the networks in the diffusion model, save the

Figure 2.5: Simulation 2, 25-agent network results: the total average proportion of innovators displayed on the y axis, and the communication round number as a measure of time on the x axis.

25-agent ones with the lowest connection probability, have either completely or nearly completely diffused the innovation by then. At transfer, each agent will have a chance to adopt the innovation equal to the proportion of agents it is connected to that have also innovated, or revert to the un-innovated state otherwise. This means that agents whose connections have all innovated will continue to innovate in the next generation, and conversely, those with no exposure to the innovation will not. The process was allowed to repeat for 100 generations.

Figure 2.6: Simulation2, 125-agent network results: the total average proportion of innovators displayed on the y axis, and the communication round number as a measure of time on the x axis.

Results & Discussion

As with Simulation 1, the results for the 25-agent and 125-agent simulations are presented in Figure 2.5 and Figure 2.6. However the results are difficult to see, even at a large print size, as many networks converge to similar proportions of innovation. As such, the average proportion of innovators in each network from generations 90 to 100 are summarized in tabular form in Figure 2.1. There they are presented in increasing order from top to bottom.

Immediately apparent, are discrete levels at which the various network topologies stabilize. This is caused by the fact that, unlike in Simulation 1, it is possible that the innovation fails to be vertically transmitted at some point and thus disappears; the innovation fails to compete against the existing conventions. In cases where the innovation does not disappear, though, it eventually goes on to overtake the entire networks. This stable level can be seen as the average proportion of networks that are able to completely innovate for the given topology.

Surprisingly, in both size conditions, almost all networks are unable to attain full diffusion; there are at least some instances when the innovation fails to spread through vertical transmission. For instance, the 125-agent network with 1.0 connection probability was able to completely diffuse its innovation by round 2 in Simulation 1. Yet here, it is only able to obtain a maximum proportion of 0.69 on average. Looking at the proportions as ranked in Figure 2.1, in both network sizes, networks with connection probabilities that yielded moderate to poor times to diffusion in Simulation 1 are now able to maintain the highest proportion of innovators.

The reason for this is less intuitive than the diffusion-only model of Simulation 1, but the mechanism of transmission brings the results closer to the expectations of the sociolinguistics literature. That is, in sparsely connected networks, it is easier for the spread of innovations to capture a large proportion of any given agent’s neighbors. Particularly, in the large networks with high clustering coefficients, an innovation must spread over almost the entire population to consistently be maintained after each generational transfer. This is the resistance effect of dense clustering to innovation described in the literature (Milroy & Milroy, 1985). Similarly, networks with low clustering coefficients have difficulty ensuring that the innovation spreads in a manner that will reproduce regularly with each transfer. This emerges from the fact that they have few triangular relationships which help spread innovations to all of an agent’s connections quickly.

Not surprisingly then, the networks with moderate clustering coefficients generally maintain the highest proportion of innovators. On average, they are more likely to contain some dense areas with triangularly connected agents, and some with more sparse connectivity. Thus there are continually pockets of density that when having adopted the innovation are able to maintain it, and less dense areas that insulate them.

Most important though, is that given a diffusion model comparable to others in current use, the addition of transmission alters the expected outcome based a priori on the diffusion model alone. Even more so, the manipulation of the clustering coefficient has a difficult to predict effect that is highly dependent on size and the actual configuration of ties in a specific network. This highlights the need to separately maintain the distinction

between borrowing of a form in diffusion, and the amount of incremental changes that have occurred in transmission.

This last point of the effect of clustering coefficients is particularly important, as there is a trend to use online communication networks as stand-ins for human social ties. Certainly it is desirable for ecological validity to try and replicate natural social networks as closely as possible. However, the networks characterized by modern, hierarchical societies tend to have similar or identical clustering coefficients, even across different sized populations (Reali et al., 2018). The relevance, then, of considering networks with different coefficients, which as a proxy for intimacy should be driving factor in change, is even more important (Trudgill, 2011). In response to such models and diffusion-centric mechanics, a novel, high-level model that embodies both diffusion and transmission mechanics will be developed in the following simulations.

2.2 High-Level Model of Language Change

In this remainder of this section, a high-level model of both the diffusion and transmission mechanisms will be presented. In particular, the modeling of transmission will capture the subprocesses responsible for the development of bound morphology as identified in the literature review. After reviewing these and explaining their implementation, first the effects of the global clustering coefficient using random networks will be tested in Simulation 3. In order to remedy problems with ecological validity at low levels of the coefficient, another set of simulations using a network specifically designed

Network	Proportion	Network	Proportion
0.1	0.43	0.2	0.59
0.8	0.60	0.8	0.66
0.2	0.63	0.9	0.66
0.7	0.66	0.6	0.66
0.6	0.73	1.0	0.69
0.9	0.73	0.3	0.69
1.0	0.73	0.5	0.69
0.3	0.76	0.4	0.69
0.4	0.83	0.7	0.72
0.5	0.86	0.1	0.83

Table 2.1: Diffusion Measure with Transmission: on the left are the 25-agent network results, and on the right, the 125-agent network results. The networks are shown in increasing order, along with the average proportion of innovators in generations 90 through 100.

to mimic the properties of modern hierarchical social structure will be run in Simulation 4. Finally, how size affects these initial results is investigated in Simulation 5.

2.2.1 Mechanisms of Language Change

As outlined in the literature review, grammaticalization originally focused on how individual words evolve into grammatical ones, but that is now seen as a specific case of more general processes of language change (Narrog & Heine, 2011). At least three interacting mechanisms appear to underly it: extension, repetition, and reanalysis². The first two mechanisms are driven

²As a reminder for those familiar with the more traditional linguistics literature, extension is often referred to as analogy, while reanalysis refers specifically to changes in grammatical structure, while an additional term, bleaching, is used to refer to changes in semantic structure. In this paper we follow a more recent trend to highlight the more general means of novel usage with the term extension (Narrog & Heine, 2011); much novelty that leads to language change has nothing to do with analogy as conventionally understood in the context of English grammar. Similarly, structural change happens at various levels of language use, and so we use the term reanalysis throughout to refer to

by competent speakers of the language, while the latter is driven by language learners.

Extension occurs when speakers use existing patterns in novel ways, often for pragmatic reasons. In the future tense example in Chapter 1, the use of the PIE desiderative to express future time resulted from the novel equating of things we desire to mean things that have yet to happen. The later PIE future with *to have* develops similarly. Such semantic pathways are often attested in other languages, and in fact, English is currently using them both: *I have to love* and *I will love* are both used to express future actions.

Repetition is the second mechanism carried out by competent speakers. As a signal becomes more predictable, it requires fewer distinguishing physical features to be understood. In the example above, as the verb *habere* became more specialized as a sign of the future tense, its occurrence became more predictable from context where future action was implied. With this increase in predictability, less articulatory effort was spent by speakers, and the length of this signal shortened as learners began to internalize the reduced realization. This loss of phonological content is often called **erosion** in the grammaticalization literature.

The final mechanism, reanalysis, contributes to bound morphology when language learners internalize such novel patterns in their input due to extension and repetition by existing speakers. While competent speakers know when a signal is being used in a novel way (extension) or is being expressed more succinctly (repetition effects), learners do not have such background knowledge of the underlying forms. Instead, there is the chance that these

structural change regardless of kind

novelties will be taken as fundamental, and the underlying language patterns learned by one generation can become subtly different from the one before. In the PIE future tense, reanalysis occurs when the PIE desiderative shifts from expressing desire for actions that have not yet taken place, to simply expressing future time without reference to desire. This, again, is precisely what happened with English *will*.

In order to create bound morphology, these mechanisms must act together: extension and repetition effects change the input to learners, who then make subtle reanalyses to the underlying structure. The gradual accrual of such small changes is what Labov (2007) highlights as the difference between diffusion and transmission, which he refers to as incrementation. In the simulation results below, the term reanalysis is used to specifically refer to these kinds of incremental reanalyses that are the result of extension and repetition, and not other kinds of structural change; that is, the reanalyses counted by the model single out the chained occurrence of mechanisms responsible for the development of bound morphology.

2.2.2 Language Model

Language is simulated as a set of meaning-signal pairings. In all experiments the agents can express one of three meanings (each represented by an integer value) that stand for a core, grammatical function, e.g. tense, aspect, number. These meanings are taken to be implicit in the cognition of events, and thus present at all stages of the language, but may vary in the signals used to represent them. Consider the constant need in the history from PIE

to Spanish to express the future time of an action, but the various signaling means employed over time.

The agents communicate with signals, represented as a sequence of four integer values. The first integer value is the *meaning* that it communicates. The second integer value is the lexical *origin* of the signal. In the PIE example, there were two lexical origins: the original lexical origin of the PIE desiderative suffix, and the verb ‘to have’ that is still present in the daughter languages. The third integer value is the number of *reanalyses* through intergenerational transfer that have occurred, and the fourth and final integer value is a unique *identifier* for the signal. As described above, a given reanalysis results in a semantic or structural change to the signal, but a signal representing the same meaning and having the same lexical origin can be reanalyzed a similar number of times with different effects. For example, Spanish and Italian share the same general historical developments of the future tense inflexion, but different sound changes took place in each set of speakers (thus there are two languages instead of one) and so even though they share the same lexical origin and similar number of reanalyses, the signals are not identical.

Agents communicate by presenting these signals to each other, and depending on their dynamic language experience, either understanding the signal or taking creative action to repair communicability. This is the diffusion mechanic of the model. After a fixed number of exchanges, each agent is replaced by a new one that acquires its linguistic knowledge from the experiences of the agent it is replacing; the new agent acquires the language from the input it receives, and forms the next generation of agents. This

is the transmission mechanic of the model, and crucially, it is at this stage that given the occurrence of extension or repetition effects, a signal might be reanalyzed.

It should be mentioned that this method of transmission is a direct instance of what is known as the **iterated Learning paradigm**, described in detail as it pertains to computational modeling in Kirby (2001). Conceptually, it is the experimental paradigm equivalent of the telephone game, where a message is passed from person to person and the changes to it through transmission are seen. It has been used in the laboratory with human participants (Smith et al., 2008) as well as computational agents (as mentioned above).

As mentioned in Chapter 1, outside of linguistics, work on language change has focused on language as the product of cultural evolutionary forces (Smith et al., 2003; Kirby, 2001). The distinction between cultural evolution and conventional, biological evolution is that each learner of a language must observe the linguistic behavior of others and infer the structure of the language from this input; one does not inherit in their DNA grammatical patterns or lexical knowledge. We do seem to inherit endogenous abilities, though, that interject a bias towards certain analyses (Han et al., 2016). As a consequence of this, small changes in the input of language learners can lead to novel patterns in one generation that were not found in the previous one.

The iterated learning paradigm operationalizes this by having the output responses of the prior participant be used as the learning input for the current participant; like the telephone game, each participant is teaching the next.

However, as in cultural evolution, each participant is not afforded a chance to see every possible input, just as in real life a person cannot hear every possible sentence in a language, or see every possible response to a situation.

While traditional uses of the iterated learning paradigm have involved one participant or one agent teaching another directly, there is no reason the outputs of many participants or agents cannot be pooled together to form a collective input. In the high-level model (and unchanged in the low-level model) each agent stores the history of the inputs it receives, and used them during replacement. In addition, it is highly unlikely that any single agent will ever see all the meaning-signal pairs in use by others in the network, and so just as in the traditional iterated learning paradigm, only a subset can be used in transmission.

2.2.3 Agents

An agent consists of four parts: an active repertoire of signals, a passive repertoire of signals, a history of usage experiences, and an identifier. The active and passive repertoires are implemented as sets of signals. The active repertoire contains signals that the agent uses to communicate with other agents, and the passive repertoire contains all of the active signals, as well as additional signals that the agent came to understand from others, but does not (yet) produce. The identifier is a unique integer value that identifies the agent's position in the network across generations, and the history is a sequence of pairs, where each pair consists of the identifier of another agent that communicated with the agent, and the signal that the other agent used

for that communication; the history is a record of who communicated what to the agent.

Initially, and after each intergenerational transfer, an agent has exactly one signal for each meaning in the language, and an empty history. At first, both the active and passive repertoires consist of just these signals, and are identical. As communication progresses, described in detail below, each agent comes to understand signals used by others and adds them to the passive repertoire. With additional exposure, signals from the passive repertoire can be added to the active, and thus become produced as well as understood.

2.2.4 Agent-Agent Communication

The communication procedure is summarized in Figure 2.7. Communication proceeds with two agents each selecting a signal from their active repertoires and presenting it to their partner. Their partner then uses their own passive repertoire to see if they know that signal already. If the signal is not known, they search the passive repertoire for a significantly close signal of the same meaning they can use to interpret it with. The mechanics of significantly close, explained shortly, capture the fact that while speakers of a language may have different internal understandings of a signal, on the surface their usage is still intelligible to others. Again, consider the future tense example. At one point a generation of Spanish speakers understood the *-é* of the future tense to be the verb ‘to have’ while their close descendants understood that same sound to be simply a verbal suffix that just expressed tense. However,

Figure 2.7: Schematic of Horizontal Communication Procedure.

Figure 2.8: Schematic of Compatible Signals

incremental changes through transmission are gradual, and on the surface all speakers were producing phonetically the same sequences, and always in a context where future time was implied. Something as seemingly different in analysis as a verb and suffix can coexist and be mutually intelligible amongst all speakers.

For one of the passive repertoire signals to be sufficiently close, and thus facilitate intelligibility, requires that the integer values of the two signals (the speaker's chosen signal and the passive one being checked) are comparable. This procedure is summarized in Figure 2.8. First, the meanings must match. Second, the lexical origins must match, so that the fundamental semantic and physical properties of the signal are similar. Finally, the number

of reanalyses must be within a set threshold from each other, thus allowing for differences to emerge between the usage of different generations, but not so different that mutual intelligibility is lost. In all simulations, that threshold is set at plus or minus one occurrence of reanalysis. Alternatively, if the signal being used has zero levels of reanalysis and thus has not undergone any of the processes that might obscure its literal meaning, it is simply a lexical construction with no potential for misunderstanding. Such zero-level signals are immediately understood without need for any specific knowledge from the passive repertoire. This behavior of understanding zero-level signals is designed to mimic the usage of adults in language contact situations. Namely, when trying to express a novel meaning, a periphrastic means of communication is typically used, avoiding potential misunderstandings due to semantic shifts or complex morphology (Hickey, 2010; Sebba, 1997).

If a signal in an agent’s passive repertoire can be found such that the signal being shared with it is intelligible, the communication is a success. The agent adds it to its the passive repertoire if not already present, and adds a pair consisting of the speaker’s identifier and the communicated signal to its history. Additionally, if the hearer already possesses the signal in their passive repertoire, i.e. it is being exposed again, it undergoes a small chance ($p = 0.25$) to add the signal to its active repertoire. In this manner, agents can adapt to and adopt the usage patterns of others, and innovations can diffuse across the network.

However, if a signal that facilitates communication cannot be found in the passive repertoire, the speaker will try and repair the communication by using any other active signals it possesses with the same meaning. Again,

for each such repair the hearer will attempt to understand it using the available passive knowledge. If no such communicable signal can be found to communicate the intended meaning, the agents will coin a new signal to fill the expressibility gap. Computationally, a new signal with the same meaning as the signal attempted to be shared is created, and given a new origin, a new identifier, and zero levels of reanalysis. This corresponds to the observed behavior in language contact and second language user situations when communication is obstructed by grammatical intricacies. In these situations speakers use periphrastic constructions, i.e. phrases in which no words contain bound morphology, that appeal to adult biases in language understanding (Hickey, 2010; Sebba, 1997). In terms of the model the new origin captures the lexical content of this new periphrastic construction, and it is given zero levels of reanalysis as it is structurally the least grammaticalized.

2.2.5 Intergenerational Transfer

The intergenerational transfer procedure is summarized in Figure 2.9. After a set number of communication events, each agent undergoes a replacement process that simulates intergenerational transfer. For each meaning in the language, the ‘child’ agent selects a single signal to represent it. To pick the representative signal, the agent searches the history of the ‘parent’ agent it is replacing and attempts to select the signal used by the most interlocutors, i.e. the signal that was most used to represent the meaning. If there is more than one candidate, then the signal is chosen at random.

After this usage-based selection, the mechanisms of extension and repeti-

Figure 2.9: Schematic of Compatible Signals

tion are simulated together, as they are linked with respect to the potential for reanalysis. Extension is a constant force in human language, as speakers continually innovate for pragmatic, rhetorical, and descriptive motivations. Similarly, as forms become more predictable, their physical structure is elided; this is the so-called erosion mentioned above. These speaker driven changes directly lead to the possibility that structural changes (whether they are semantic, grammatical, or phonological) occur. To account for this, each one of the selected signals undergoes reanalysis as determined by a small base probability for change ($p = 0.01$), representing latent extension and repetition that all speakers interject. This small chance is then multiplied by the number of unique speakers that used the signal in the history. This chance is calculated for each signal that was chosen in the previous step, and each signal probabilistically increases its level of reanalysis based on its individual calculated likelihood. If a signal does undergo a reanalysis, its reanalysis level is increased by one, and it is assigned the next available identifier to capture the fact it has changed. Again, it should be noted that our model does not attempt to distinguish the nuances of different kinds of structural changes (e.g. semantic bleaching), but rather the more general fact that structural change has occurred in response to extension or repetition.

Finally, signals can eventually become so reduced in physical form that they become eroded from the language as learners fail to detect their presence. Recall the gradual reduction from the full Vulgar Latin verb *habere* to *-é* in Modern Spanish. Eventually, signals are so reduced that they become lost, much like the PIE desiderative suffix that disappeared and created the need to coin the periphrastic construction with *habere* in the first place. To

account for this, any signal that is above a set threshold for reanalyses (level greater than 10) undergoes a chance ($p = 0.3$) to be lost from the language. If a signal is lost, the speaker replaces it with a new periphrastic signal, using the same procedure as when two speakers cannot communicate.

Once the new signals have been selected and undergone the mechanisms of language change and loss, the new agent is ready to communicate. It begins with an empty history and with active and passive repertoires that consist of just these chosen signals.

The principle behind the setting of these parameter is largely based on intuition, but some that are suspected to directly affect the results are explored at the end of Simulation 5. In particular, round length and probability for change are explored there, and will not be mentioned here. Among those that remain, the limit threshold of 10 reanalysis was based on the observation that it seems on average about 5 structural changes occur in a large collection of grammaticalization histories (Heine & Kuteva, 2002) and about 2 phonological changes before a structural change (Narrog & Heine, 2011). This is not an strict boundary by any means, and so a small (one third) chance that a meaning-signal pair will be eliminated above this threshold was added. Should this threshold be increased, the model would be able to retain meaning-signal pairs for longer, and thus there may be more potential room to see the absp;ite effects of the networks.

Similarly, the one in four chance a meaning-signal pair will be adopted is based on the observation it takes some exposure to acquire a novel form; we don't immediately repeat the novel expressions we encounter in life. As was noted in Real et al. (2014), it is indeed harder for forms with more

bound morphology to diffuse than lexical, and so forms with many reanalyses might be more difficult to diffuse than not. However, as the reanalysis level is not an actual measure of bound morphology, this dynamic aspect of diffusion probability was not included. Regardless, the probability of adoption directly controls the rate at which meaning-signal pairs diffuse, as when higher, fewer exposures are required for it to become actively spread by a potential adopter.

2.3 Simulation 3: Network transitivity and linguistic reanalysis

As mentioned in the literature review linguists have proposed the ‘intimacy’ of small communities as a crucial factor for supporting high levels of synthesis (Nettle, 2012; Trudgill, 2011). The idea behind intimacy is not that just more variation happens, but that when novel usages occur, they are a much more substantial portion of a learner’s input. Rather than being a one-off oddity of speech, the continual contact with the same people in a small, intimate society allows for sustained exposure to these variations, and thus increasing the pressure to account for this with reanalysis during acquisition.

Indeed, in a small intimate society one is hearing the same things over and over again. However, when one of these novel forms is repeated over and over again, it provides an affordance for reanalysis in the system, and is much more likely to take place. In turn, the density of ties allow for quick diffusion of innovations within the community, increasing the likelihood they

become adopted and available for repeated reanalysis in the future.

The goal of Simulation 3 is to capture the idea of intimacy mechanistically and in a quantifiable way. As discussed in the literature review on the clustering coefficient, it directly measures the kind of triangular ties that emerge in societies of intimates, and so is used here as a proxy for the level of intimacy in a society. As a reminder, the clustering coefficient can be measured in two ways: globally, reducing the entire network to a single density measure, or locally, capturing the density surrounding an individual agent. The global measure is defined as the number of closed triplets divided by the total number of triplets. A triplet is a unique set of three agents and two connections that link them together in the graph. A closed triplet is a unique set of three agents, and three edges that link them together, i.e. a triangular relationship amongst the agents. The local measure of transitivity, defined for an individual agent, is the number of edges that exist between the agent’s neighbors divided by the total number of possible edges that could exist between the neighbors. Again, the two terms, transitivity and clustering coefficient, will be used interchangeably, but in particular the term transitivity will be favored in figures to conserve space.

2.3.1 Network Architecture and Communication

To assess the effect of the clustering coefficient, *random networks* (Erdős & Rényi, 1959) were constructed with varying probabilities of connection between agents. Connection probability in random networks is approximately equal to the global clustering coefficient (Barabási, 2014), and so a spec-

Parameter	Value
Reanalysis Threshold	+ - 1
Adoption Probability	p=0.25
Erosion Threshold	20
Erosion Probability	p=0.33
Variation Probability	p=0.01

Figure 2.10: Example of Random networks used in Simulation 1, with connection probability 0.2 (top) and connection probability 0.7 (bottom). Model parameters are listed to the right.

trum of networks was created with connection probabilities ranging from 1.0 to 0.1, with intermediate step sizes of ten percent. Figure 2.10 shows two random networks with connection probabilities of 0.7 and 0.2 among a population of 25 agents. These are the same kinds of networks, created with the same range of connection probabilities as in Simulations 1 and 2 earlier.

Each network was constructed with 25 agents, and used the same initial language across all simulations. Each agent was initialized with the same signal set: one signal for each of 3 possible meanings, where each signal possessed no initial level of reanalysis; all agents initially used the same, completely periphrastic signals for communication. For each connection probability level, 100 unique networks were generated and allowed to undergo 1000 intergenerational transfer events. Before each intergenerational transfer, agents were allowed to communicate for 10 rounds of communication,

where a round of communication consisted of each agent communicating a randomly selected signal from their active repertoire to each of their interlocutors in the graph. The order of these communications was randomized in each round to avoid ordering effects.

The communication and intergenerational transfers were conducted as explained in the language model section above, with the specific parameters as given in Figure 2.10: the reanalysis threshold is how close the reanalysis levels of two signals, otherwise matching in meaning and origin, could be and still be considered intelligible. The adoption probability is the chance at each repeated exposure to a known signal that it would be promoted from passive to active use. The erosion threshold is the maximum reanalysis level a signal could reach, above which, each time the signal went through intergenerational transfer it was removed with chance equal to the erosion probability. Finally, variation probability is the level of latent extension and repetition of the agents; the number of unique speakers experienced using the signal, multiplied by variation probability, yielded the chance that a signal would be reanalyzed.

2.3.2 Prediction

As an agent’s learning algorithm is usage-based, and sensitive to the kind of signal variation described in the linguistic literature, it is expected that network capacity to support sustained reanalysis will correlate positively with increasing levels of the clustering; more intimacy, as operationalized through the clustering coefficient, should predict higher levels of reanalysis.

This should happen for two reasons. First, high levels of clustering result in greater variation in each agent’s input, and second, they provide denser connectivity that aids dispersal of innovations. Finally, as the statistical correlation of social demographics and morphological complexity was linear in the WALS database (Lupyan & Dale, 2010), a similar linear relationship was expected.

2.3.3 Results

In order to track the capacity of a network to support the mechanisms responsible for developing bound morphology, the levels of reanalysis present were measured after each intergenerational transfer. At this point in time, directly after intergenerational transfer, every agent has the same number of signals - one for each meaning. The agents have acquired their individual signals from experience, but have yet to begin communicating. The reanalysis level of every signal used by every agent in a given network was totaled, and then divided by the number of agents, to provide a mean value estimate for the reanalysis of a typical signal of the language. The mean level of reanalysis is displayed in Figure 2.11 with each network colored by measure of global clustering coefficient, also called transitivity. After initial fluctuations in reanalysis values, all networks stabilized by the 500th generation. A mixed-effects model was fitted with individual networks as a random effect, and the global clustering coefficient, as a fixed effect, to predict the mean levels of reanalysis as the dependent variable. Data was selected from the last 1000th generation, as only the stable behavior of the models at the end

Figure 2.11: Results of Simulation 3: mean levels of linguistic reanalysis as a function of network connectivity over 1000 generations.

of their evolutionary path was of interest. In line with the prediction, the statistical model yielded a positive effect of the clustering coefficient ($\beta = 3.27$; $t=9.36$; $p < 0.001$), with more connected networks sustaining higher values of reanalysis. The details of this analysis are found in the appendix of R reports under the same section heading.

While the interpretation of the 10 levels of reanalysis the networks with clustering above 0.4 are attracted is discussed below, it likely has a simple empirical explanation in the model parameters. The threshold before meaning-signal pairs are able to drop was set at 10 level of reanalysis, and with a 1/3 chance at each successive transmission that the drop occurs. It is likely, that a ceiling affect is seen here, where the networks of 0.4 may potentially be able to support higher levels if the threshold was increased.

Figure 2.12: Local Transitivity: mean levels of linguistic reanalysis as a function of local transitivity. The top two graphs and bottom left are averaged over the individual generation listed, and the bottom right is averaged over all generations. The blue line is the LOESS fit provide by the ggplot2 package in R

2.3.4 Discussion

As a physical proxy for ‘intimacy’ the effect of the clustering coefficient is in line with correlational findings that more intimate societies develop greater usage of bound morphology (Lupyan & Dale, 2010). As network clustering increases so does the capacity for repeated linguistic reanalysis, and thus the likelihood of bound morphology being developed. However, the effect of clustering on linguistic reanalysis is non-linear, evident in the unequal increases in reanalysis despite equal increase in transitivity across conditions, similar to the effect seen in the diffusion model of Simulation 1. Further, there appears to be a threshold between 0.3 and 0.4 for which networks above the threshold perform qualitatively identical. Above the

threshold, reanalysis quickly rises and reaches a high stable state, while in contrast, below the threshold reanalysis stabilizes at far lower levels.

To explore the mechanism driving this finding, the corresponding agent-level measure of clustering, the local clustering coefficient (or local transitivity), was examined. The mean level of reanalysis as a function of local transitivity is presented in Figure 2.12. Each data point is averaged across all networks and agents in Simulation 3 with the given local transitivity measurement.

As seen after one intergenerational transfer (top left) there is virtually no effect of the local density of connections surrounding an agent. However, by generation 100 (top right) the interconnectedness of an agent's immediate neighbors is highly correlated with the level of reanalysis in the signals it produces. By the end of the networks' simulation at generation 1000 the reanalysis levels have stabilized, as seen in Figure 2.11 above. Looking at the local transitivity measures for generation 1000 (Figure 2.12, bottom left), though, we can see that there is still substantial variation between agents, and increasingly so as local transitivity declines. Indeed, averaging over all generations (bottom right) agents with low local transitivity produce signals with lower levels of reanalysis across the lifetime of the network.

It appears that the differences in global levels of reanalysis observed in Figure 2.11 are connected to the local levels plotted in Figure 2.12. The distribution of local transitivity by network group is plotted in Figure 2.13, with the top row consisting of the two networks immediately above the threshold and the bottom row the two networks immediately below the threshold. In networks below the threshold (bottom row), the majority of agents pos-

Figure 2.13: Distribution of Local Transitivity: agent counts are taken from all networks generated at the listed connection probability.

sess local transitivity values that place them in the highly varied region for average reanalysis as seen in Figure 2.12. In this region the range of different reanalysis levels among the agents is greater than the threshold for communicability, which causes periphrastic signals to be generated in order to repair this communication gap. Conversely, those networks with global transitivity above the threshold have the majority of their agents with local transitivity values along the stable, constant region for averaged reanalysis.

While the networks with high levels of transitivity are good models of small communities, those with low levels are not. Human social networks, are characterized by dense connectivity, especially in small groups (Newman, 2010), and so the low transitivity networks in Simulation 3 are poor proxies for reality. While the random network models used in this simulation allow the effect of transitivity to be easily observed in a controlled fashion, the language model must be applied to networks that accurately embody the network properties of both large and small human societies.

2.4 Simulation 4: Network Topology and Linguistic Reanalysis

As mentioned in the literature review, the web of social interaction in small communities is dense, with most members communicating with most other members. As societies grow, though, they lose this dense connectivity, and communities within communities form (Newman, 2010). People have limits on the amount of time and effort they can input into social interactions

(Dunbar, 1998; McCarty et al., 2001), and as population size increases, this pressure results in local clusters of social ties (Jin et al., 2001). In large linguistic groups, e.g. English speakers, people socialize with a small fraction of the whole. Human social networks, both large and small, demonstrate high levels of clustering, but large communities display a major structural characteristic that small communities do not: large networks are typically scale-free (Newman, 2003; Ravasz & Barabási, 2003).

A scale-free network, technically, is one in which the degree distribution follows a power law (Barabási, 2014). In terms of social structure, the effect of this degree distribution is that on average most agents know only a small percentage of the others, while there exist some agents that know a much larger percentage. These so-called ‘hubs’ agents that have numerous connections serve to interconnect communities. It should be noted that for the mechanism discussed here, only the fact that hubs emerge in connection to scale freedom is important.

2.4.1 Network Architecture and Communication

Four distinct network topologies were modeled. In order to capture the connectivity patterns of small groups, a *complete network* (Complete) was used, where by definition all agents interact with all others; it is analogous to the organization found in a small village. This is the top network of Figure 2.14. In contrast, to capture the topology of larger societies, a *hierarchical network* (Hierarchical) designed specifically to mimic the social hierarchy of larger human social networks was used (Ravasz & Barabási,

Figure 2.14: The Complete (top left) and Hierarchical (bottom left) networks used in Simulation 4 exemplify connectivity properties of two common types of human social communities respectively: the village and the modern city. The transitivity measures for all networks (BA and Random not pictured) are given to the right (the Complete network depicts only 10 agents for clarity).

2003); this is a model of connected communities, analogous to a physically divided network of villages or the socially divided network of a modern city. It is shown in the bottom of Figure 2.14. By way of control, two additional networks were tested. The first, *Barabasi-Albert* (BA), is commonly used in diffusion models of innovation (Ke & Yao, 2008; ?), and allows for the creation of a scale-free network with zero transitivity (Barabási & Albert, 1999). By eliminating transitivity from the network, the effect of hub agents can be assessed in isolation. In diffusion models the presence of hubs drive the spread of innovation by linking large numbers of agents that otherwise may be distant from each other.

It would be ideal to also compare the common BA model with the clus-

tering coefficient present in the hierarchical network; however, a number of problems exist for this at present. First, the traditional BA model does not produce the requisite levels of transitivity present in observed social networks (Ravasz & Barabási, 2003; Varga, 2015). Second, while modified versions of the BA model exist to try and make this property an adjustable parameter (Jin et al., 2001; Varga, 2015), as well as novel procedures to produce scale-free networks with tunable transitivity (Chakrabarti et al., 2017; Herrera & Zufiria, 2011), no method yet exists that accounts for the distribution of connection density in a way that mimics social networks, and/or allows for manipulation of that density measure beyond a narrow range. A key issue is that in large, scale-free human social networks, the local transitivity of an agent is inversely proportional to its number of connections, i.e. to its degree (Soffer & Vazquez, 2005). At present, generative models of social network structures are an ongoing area of active research, as represented in the above works.

Lastly, a *random network* (Random) was also created (Erdős & Rényi, 1959). The connection probability was $\frac{\ln N}{N}$, the threshold for single connected components, and where N is the number of agents. This allows for the creation of a non-scale-free network with no assumptions about the hierarchical distribution of agent ties.

As with Simulation 3, each network contained a fixed number of 25 agents across all conditions, starting with the same initial language at generation 0. Only the topology was varied. The population size of the agents, the initial language, and the number of communications per agent were the same in all conditions; the only difference being the patterns of connectivity. All

simulation parameters were identical to Simulation 3, Figure 2.10. As with Simulation 3, each network topology was allowed to run for 1000 generations of transmission, and each topology was run for 100 replication.

2.4.2 Prediction

The transitivity values for these networks are summarized on the right hand side of Figure 2.14, and both Complete and Hierarchical are above the threshold for maximal linguistic reanalysis seen in Simulation 1, while BA and Random are far below it. If it is the case that transitivity alone affects the ability of a network to support sustained reanalysis, then given these values Complete and Hierarchical are expected to support similar, maximal levels of reanalysis, and BA and Random to support little to none.

However, Simulation 3 did not account for the kinds of hierarchical structure seen in actual human social networks, it was simply an investigation of the effects of the clustering coefficient on the model. If instead it is the case that the hub connections that develop in scale-free networks are important for supporting sustained reanalysis, then we would expect BA and Hierarchical to have an advantage over Complete, as both of these networks are scale-free while Complete is not. This is especially so for Hierarchical, which has a transitivity level above the observed threshold for maximal reanalysis observed in Simulation 3.

Figure 2.15: Simulation 4, 25-agent network results: mean levels of linguistic reanalysis in four different 25-agent network architectures for 1000 generations. 100 replications per network type.

Figure 2.16: Simulation 5, 125-agent network results: mean levels of linguistic reanalysis in two different 125-agent network architectures for 500 generations. 100 replications per network type.

2.4.3 Results

Complexity was measured using the same method as Simulation 3, and is plotted in Figure 2.15 as a function of generation and network topology. Different paths of evolution of reanalysis are noticeable across network topologies. In particular, Complete exhibits an initial large increase with further stabilization, while Hierarchical and Random proceed along more gradual increases, while Barabasi hardly develops. To confirm these trends, a mixed-effects model was fitted with individual networks as random factor, and topology at the final generation as a fixed factor, to predict amount of reanalysis. The model yielded main effects of topology, such that the Complete yielded larger levels of reanalysis. In particular, as predicted the Complete network developed significantly higher reanalysis than the Hierarchical model ($\beta = 4.27$; $t=60.762$; $p<.001$). Conversely, the Barabasi model exhibited significantly lower reanalysis than the Hierarchical model ($\beta = -5.09$; $t=-72.42$; $p<.001$). The Random network was statistically different from Hierarchical ($\beta = -0.21$; $t=-3.87$; $p=0.01$), but in comparison had a negligible beta value. The complete analysis formula and R report can be found in the R reports appendix.

2.4.4 Discussion

A seemingly surprising finding is that despite the level of transitivity being above the threshold for maximal reanalysis, the capacity for reanalysis in Hierarchical is not practically different than Random. A crucial difference, though, is in how that clustering is distributed within the networks. The

Figure 2.17: Study of Hierarchical Network 3: top figure displays average reanalysis levels across all 100 generations of the network. The middle and bottom figures center around the large drop in reanalysis at generation 500, and are broken down into the physical clusters of agents (middle) and central hub agent vs. others (bottom).

high transitivity in Hierarchical is due to the dense connectivity within its five-agent clusters, c.f. Figure 2.14, with a transitivity of 0.7 each. However, communication between these small, near-complete communities is mediated through the central agent of the central cluster, the *hub* agent of the network. It is this mediation through the hub that counteracts the complexity generated by these dense clusters.

Zooming in on a single network provides a clearer interpretation. The top of Figure 2.17 plots the average level of reanalysis over the 1000 generations of a single, typical run of Hierarchical. Characteristic of simulations of Hierarchical is the steady climb of reanalysis levels, as predicted by the high transitivity of the individual clusters, but punctuated by both small and catastrophic declines. While there is always the possibility that a mutually

intelligible signal with a lower level of reanalysis may replace another, the small and catastrophic declines are caused when the signals in one group become mutually unintelligible with the others. It should be noted that most hierarchical networks show multiple catastrophic declines, and this network singled out as being a rare instance of only having one. While no simulations were allowed to run longer than 1000 generations, as explained shortly, the presence of hub agents would likely continue to cause such declines indefinitely.

It is impossible for the hub agent to learn and actively produce every distinct signal for all five clusters, and thus over many generations the signals within a group may become locally more complex than those of their neighbors, or become so complex that they erode and are replaced. At this point communication is broken across the groups, and periphrastic signals are generated as the hub tries to maintain communicability. These periphrastic signals, which are automatically understood regardless of passive knowledge, may spread and replace the natural grammaticalization process of another origin. This would be analogous to the *will* future in English becoming dominant and replacing the more reanalyzed *going to* future and its variants, e.g. *gonna*, *gun'*, *goin'* etc. It should be emphasized again that while such replacement as *going* with *gonna* indeed has shown a two morpheme form being reduced to a one-morpheme form, the complexity of the English grammatical system was not changed; both forms exist as alternatives, much like competing meaning-signal pairs in the model.

Often, this small reduction in average reanalysis is isolated to the central community in which the hub is embedded and one or two local clusters

with which communication was repaired, and only results in a small dip in overall network reanalysis levels. However, the hub is in contact with all five communities, and even when its own central cluster and others are undergoing such a replacement, it may continue to acquire signals with high levels or reanalysis still in use in other groups. This distinct hub behavior leads to catastrophic declines, as occurs around generation 500 in Figure 2.17.

The bottom two graphs of Figure 2.17 are centered around this catastrophic decline, with the middle graph showing the average complexity in the individual clusters (named for their physical positions in Figure 2.14) and the average complexity of the hub agent vs. the others. We can see that a repair happens between the central and bottom left groups, causing the sharp decline in average reanalysis as periphrastic signals replace existing ones. However, even as the hub's central group drops ever lower, the hub-agent itself retains high-level signals until all but one group has declined. At this point, the hub agent has caused a number of now competing periphrastic variants to emerge, and rather than quickly recovering, the entire network must undergo a period of little to no growth; compare the quick v-shaped drop and rise after small declines, such as at generations 170 and 210, with the protracted and u-shaped pattern surrounding the catastrophic decline around generation 500. Not until winning variants emerge can they slowly undergo joint reanalysis in the individual clusters.

This innovative behavior of hubs is also predicted by cultural evolutionary theory, as they form **bottlenecks** that restrict the flow of information (Smith et al., 2003). When the transfer of social knowledge is incomplete,

there is a pressure exerted to innovate and fill the same cultural and communicative function. The hub agents do precisely this, responding to the pressure caused by gaps in communicative ability, and innovate new signals to fill them. If this constraint to maintain communicability were removed, the five local clusters in Hierarchical would go on to develop individual languages, and the linguistic community would fracture.

2.4.5 Exploring Parameters

Finally, the periodicity of these collapses was investigated with additional simulations that examined the interaction of two key model parameters: the number of communication rounds before intergenerational transfer, and the multiplier used to determine when an innovation that leads to reanalysis occurs.

One would expect that increasing the rounds of communication between transmissions would reduce the frequency of collapses like those seen in Fig. 2.17. The chance that two agents are unable to communicate decreases, as each additional round of communication provides more opportunities for innovative signals to diffuse across the network. If agents are more readily able to maintain communication then fewer repairs will take place, and thus fewer opportunities to trigger a collapse will be present.

Conversely, as the probability that innovations are introduced increases, one would expect that the number of collapses would also increase. As more innovations occur, there are more signals in competition for expressing a given meaning, and this competition reduces the likelihood that agents are

Figure 2.18: Interplay of Parameters. On the x-axis are the number of rounds of communication allowed before transmission, and the color codes for values of the variation probability, the multiplier used to determine when innovations lead to reanalyses. The y-axis is the mean number of generations between collapses.

able to communicate. With each failure to communicate, a repair is made that may potentially lead to a collapse.

To explore this interplay, the effects of these two parameters were tested with the following values: rounds of communication was varied across 10, 20, and 30 rounds, while the variation probability was tested at 0.001, 0.005, 0.01, 0.05, and 0.01. As a reminder, the simulations reported in the previous sections had 10 rounds of communication before transmission, and as reported in Fig. 2.10 (right) a ‘variation probability’ of $p=0.01$ was used.

In order to calculate the average time between collapses, local maxima and minima were identified using the discrete analogs of traditional calculus. Specifically, the forward difference between points was used to approximate the first and second derivative information. As seen in Fig. 2.17 (top) the

mean reanalysis per generation is quite noisy, with many small rises and falls, and so to de-noise the data only every tenth generation was considered. On this smoothed data, any time a drop in mean reanalysis of greater than 50 percent occurred, it was classified as a collapse. For each pair of conditions, and additional 100 hierarchical networks were run, and the results summarized in Fig. 2.18.

As predicted for variation probability, increasing the likelihood for additional competing innovations has the effect of decreasing the average time between collapses. However, in general, the predicted effect of increased communication rounds is not present. That is, as more rounds are afforded to the agents to communicate, the average time between collapses remains constant. The lowest variation probability ($p=0.001$), though, does conform to the expectation that increasing communication rounds also increases stability.

It seems then that the stability of such hierarchical networks is most sensitive to the general amount of innovation introduced by the speakers. Moreover, this constraint on stability seems to persist despite increased diffusion of signals across the network. However, it may be the case that if innovation is so gradual as to not cause competition, increased diffusion does translate to increased stability. In particular, for the $p=0.001$ case, at 30 rounds of communication before transfer, the average interval between collapses is approaching the length of the simulation. This means that, behaviorally, as communication rounds increase the collapses that characterize the hierarchical networks (as opposed to the complete networks) may be averted.

2.5 Simulation 5: Revisiting Network Size

In Simulations 3 and 4, only the connectivity patterns of the agents were being manipulated - all networks consisted of 25 agents throughout, and began with identical languages. While population size has been reported recently as a highly predictive variable in other computational and correlative studies of language complexity (Reali et al., 2018; Bromham et al., 2015), it is also conflated with social structure. In particular, as societies grow the local clustering of agents becomes inversely proportional to its number of connections, i.e. to its degree (Carneiro, 1967; Soffer & Vazquez, 2005). In addition there is evidence that people can only maintain a finite number of active relationships (Dunbar, 1998), which may limit the range of possible structures for individual agents. In particular, the size of a complete or near complete networks may be limited to about 100-200 social connections (Gonçalves et al., 2011). In order to examine how our language model may be informed by these findings, 125-agent versions of Complete and Hierarchical used in Simulation 4 were constructed. In addition, the random networks from Simulation 3 were also run with 125-agent counterparts.

As described in Ravasz & Barabási (2003) the structure of the Hierarchical network has a fractal like structure. If we start with a 5-agent cluster, 5 of which are seen in the 25-agent model, we have 1 agent in the middle fully connected to each of the 4 others. This agent in the middle is like the central hub agent in the 25-agent version. Now if we replace each of those 5 agents with a copy of the entire 5-agent network, we would end up with the 5 clusters of the 25-agent version. If we connect the 4 periphery clus-

ters to the central hub agent, then we have produced the 25-agent network. Similarly, if we take each of the 5 clusters in the 25-agent network, and replace them with the a copy of the entire 25-agent network, we would have a 125-agent network composed of 5, 25-agent networks. If we take each of the agents peripheral agents in the individual 25-agent clusters, and connect them to the central agent of the central 25-agent cluster, we have produced the 125-agent version of the Hierarchical network.

To keep computational time to run on PC computers, the replications of each network topology were reduced from 100 to 50. In addition, because network behaviors in Simulation 5 stabilize after around 300 generations, the number of generations was reduced from 1000 to 500. Finally, as BA failed to sustain any levels of repeated reanalysis, and Random has a connection probability very close the 0.2 network, these were excluded from the 125-agent high-level, and all low-level model runs.

2.5.1 Results and Discussion

A mixed-effects model was fitted with individual networks as random effect, topology and network size, and their interactions as fixed effects. The dependent variable was the mean level of reanalysis for each network at generation 500, as seen in Figure 2.16. The model yielded main effects of topology, such that the Complete yielded larger levels of reanalysis ($\beta = 3.65$; $t=10.79$; $p<.001$). A main effect of size also obtained, with 25-agent networks yielding overall larger values of reanalysis over 125-agent networks ($\beta = -5.41$; $t=-12.88$; $p<.001$). Finally, the topology x size interaction was

significant ($\beta = 4.51$; $t=7.71$; $p<0.001$). As with the previous simulations, the full R statistical report is available in the appendix under the appropriate section.

Figure 2.20 reveals that driving the interaction was a larger mean difference in reanalysis between Hierarchical and Complete for the 125-agents. In addition, mean values of reanalysis for Complete were comparable, if not higher in the 25-agent compared to the 125-agent networks, while mean values of reanalysis were lower for Hierarchical in 125-agent networks. This pattern of results suggests that network size amplifies the effect of network structure not by boosting reanalysis levels in larger density networks, but by limiting the opportunities for sustained reanalysis in larger Hierarchical networks.

While the relative performance of these networks at the 125-agent level mirrors their 25-agent counterparts, their trajectory before reaching stability was different. In the 125-agent Complete, signals rose in reanalysis in virtual unison, until becoming so high-level that then began to erode and be replaced. In the 25-agent Complete, the signals did not all rise and erode in such a tight time scale, and so there are some perturbations before the steady state is reached. Looking at the 125-agent Hierarchical, it undergoes a slow rise like its 25-agent counterpart, and peaks as each of its five 25-agent Hierarchical clusters stabilize. However, these clusters themselves participate in a hierarchical arrangement, and the presence of these additional hubs leads to additional declines, driving the overall complexity down again.

Finally, in Figure 2.19 we have the results from the 125-agent runs of

Figure 2.19: Simulation 5, 125-agent random network results: mean levels of linguistic reanalysis as a function of network connectivity over 500 generations.

Figure 2.20: Simulation 5: interaction effect of topology x size for Complete and Hierarchical.

the 10 random networks from Simulation 3. A mixed-effects model was fitted with individual networks as random effect, and the global clustering coefficient, as a fixed effect, to predict levels of reanalysis. Data was selected from the last generation, as only the stable behavior of the models at the end of their evolutionary path was of interest. In line with the prediction, the statistical model yielded a positive effect of the clustering coefficient ($\beta = 8.79$; $t=30.73$; $p<.001$), with more connected networks sustaining higher values of reanalysis.

Similar to the effect of size on the simulations discussed in the first half of this experiment, size seems to amplify the effect of the clustering coefficient. For networks with low levels of clustering, specifically those below the threshold seen in the 25-agent simulations, their lack ability to support reanalysis is quite similar. It seems that the hindering effects of not having enough triangular relationships has been amplified to the point where they are almost identical in level of support. Relatedly, the gradation between networks above the threshold have been amplified to support the extremely high levels of reanalysis seen in the 1.0 network. The most interesting result, though, is the 0.4 network that sits just slightly above the threshold seen in Simulation 3. While its transitivity is low, it able to slowly climb higher with each generation. Due to the unstable nature of the qualitative results, additional generations were run to confirm.

As we can see in Figure 2.21 and additional 500 generations of run time (1000 in total) does indeed show that the 0.4 network (forest green) does catch up with the networks with higher clustering. While in the graph it actually surpasses the stable value, it is likely that this is similar to the

Figure 2.21: Simulation 5, 125-agent random network results: mean levels of linguistic reanalysis as a function of network connectivity over 500 generations.

other networks performance in first surpassing and then settling down to the mean value of about 10 reanalyses on average. For example compare the trajectory of the 0.5 network in aqua green.

2.6 Conclusion

It has long been posited that dense, ‘intimate’ social ties are responsible for creating conditions that support greater morphological complexity. In so far as network transitivity captures that idea, the ability of networks to sustain reanalysis of signals over many generations supports this hypothesis. The global clustering of a network is a reliable predictor for the performance of complete and near complete social networks, which model small and localized populations of speakers. Further, the effect of global clustering could be seen

to reflect different behaviors in the individual neighborhoods of agents, as captured by the measure of local clustering.

However, as seen in Simulations 4 and 5, the capacity for dense connections in a network to support reanalysis at high levels, can be largely overridden as hierarchical structure is introduced. This is contrary to what diffusion models would predict (Ke et al., 2008; Niyogi & Berwick, 1997; Baxter et al., 2006), and underscores the need for modeling in network science to closely consider means of propagation, as well as looking beyond questions of ‘time to spread’ or ‘time to convergence’ when considering intergenerational phenomena.

Beyond positing a possible mechanism underlying the observed correlation between morphological complexity of a language and the demography of its speakers, this research may have practical implications as well. Language maintenance and revitalization efforts are increasingly important as languages spoken by smaller languages continue to be lost as globalization prioritizes larger languages of economic and political importance. In particular, many such small languages possess high degrees of complexity which may be best retained if minimum levels of local transitivity are ensured, and hierarchical structure in the domains where revitalization efforts take place are minimized.

The present work represents an initial exploration at the intersection of network science and computational linguistic modeling. For instance, the model generalizes over the distinctions between semantic and grammatical structure, instead focusing on the general process that unifies them as they relate to bound morphology. It may very well be the case that further more

complex interactions emerge once this level of resolution is added in future work. Also, as clear from the contrast of the hub versus other low local-transitivity nodes in the random networks, the effects of network structure surrounding such measures is also clearly an area requiring further effort.

Further afield, while the simulations focused on the intimacy of speakers as a causal variable, it is entirely possible that structural aspects of natural languages be affected by a combination of other social variables. Some of these are discussed at length by Trudgill (2011) and Nettle (2012): Community size and degree of isolation; degree of language contact; type of language contact; degree of social stability; density of social networks (modeled here); amount of communally shared information; vocabulary size of the community; ratio of child to adult language learners in the community; and so on.

The findings match the observations made by linguists, as well as the statistical findings derived from the World Atlas of Language Structures on population spread and compositional structure (Lupyan & Dale, 2010; Dryer & Haspelmath, 2013). Small populations with high clustering coefficients are able to support sustained reanalysis, and thus the conditions for potentially morphological complexity exist. Conversely, the effect of hierarchical social structure, which emerges as human social networks grow, places a pressure for innovations that ease communication rather than being the product of repeated reanalysis. In these hierarchical structures. This complex interaction between network structure and usage-based transmission provides the first hints to a mechanistic explanation consistent with both linguistic theory and observed history of natural language change.

However, this high-level model is not directly measuring synthesis, only a proxy for it - the capacity of the network to maintain sustained chains of extension/repetition effects, followed by reanalysis. Moreover, those processes themselves were simply implemented as a probabilistic function of network structure and signal usage. In order to investigate more deeply the connection between the clustering coefficient of a social network and bound morphology, a low-level model must be developed in which morphological structures can physically emerge, and synthesis measured directly.

Chapter 3

Low-Level Implementation

In this chapter a low-level model of morphological complexity is presented. In the previous, high-level model of the last chapter, the key mechanisms of grammaticalization were captured using probabilities. More-over, the structure of the model was abstract. For example, extension was modeled as a base probability, multiplied by the number of unique individuals that the agent experienced using a particular form. In this chapter, compositional structures are used to define the meaning and signal spaces, and the fundamental processes of extension, reanalysis, and repetition operate directly on these forms. Before we begin, it should be noted that the low-level model presented herein does not attempt to model phonology, syntax, or morphology in a form approaching the complexity of what is found in natural language. For instance there are no phonological features, pragmatics, or mildly-context sensitive derivations. These things certainly could (and at some point in the future, should) be modeled, but for reasons of computational complexity and scope, they are not done in this thesis.

In addition, while the low-level model allows for a structural distinction between morphemes and words (explained in more detail below), it does not make distinctions between different kinds of word or morpheme classes. For instance, the computational rules learned in the model are unable to capture the distinction between inflectional vs. derivational morphology, nor is the semantic space structured such that nouns are distinguishable from verbs. These features can be added to the model, but presuppose a level of detail beyond the scope of directly modeling the emergence of bound morphology. For example, as discussed in Chapter 1 in the section on grammaticalization, regardless of grammatical vs. lexical content, the process by which free words become bound appears to be the same (Narrog & Heine, 2011). Similarly, the semantic content that distinguishes nouns vs. verbs, do not appear to be affected differently by the processes that support reanalysis; the visual properties of an object are just as capable of participating in an analogy as the motion properties of a verb.

As with the high-level models in the previous chapters, what follows here is a detailed prose description of the model and mechanics. For those interested in more computational details, the pseudo-code is available for the algorithms in an appendix. For those wishing to reproduce the results or borrow code, a heavily commented version of the Julia language source code is available, in addition to the R scripts used to conduct the analysis.

It should be noted that the model presented here does not follow the general trends of Natural Language Processing (NLP), especially in regards to the treatment of grammar induction. State-of-the-art approaches, as summarized in Martin & Jurafsky (2009), almost exclusively represent meaning

as a integer valued vector in a n -dimensional space. This approach was popularized widely in the Word2Vec model of Mikolov et al. (2013) adopted by Google. With regard to language evolution, work has been done using vector based models (Brighton et al., 2005; Brighton & Kirby, 2001). However, as described in Brighton et al. (2005), it is impossible to make the distinction between words and morphemes, and instead, compositionality is estimated by comparing changes in signal length and meaning vector distance. That is, if two signals that are similar differ by a single character (distance is small), but their meaning vectors are far apart in the n -dimensional space (distance is great), it is implied that the relationship between the two meaning-signal pairs is a compositional. By comparing all the meaning-signal pairs in the repertoire used by the agents, one can measure the avg change in signal distances vs. the avg change in meaning distances, and reason about the relative levels of compositionality in the language.

This work has not taken that approach for three key reasons. Firstly, it is impossible to tell exactly which changes in signal contribute to which changes in meaning; one can detect the magnitude of the change, but not its details. Secondly, there is no notion of hierarchical structure in the meaning space, meanings are points in an n -dimensional space, which as will be discussed in the following section, is a major component of semantic analysis. Finally, linguists have long developed theories that deal in features (e.g. phonological features), and the relationship between collections of features and linear algebraic spaces is not intuitive, if even possible. For these reasons, a more old-fashioned approach is used in order to capture these deficits, specifically one that allows for measuring not only the actual compositional structure of

the meaning-signal pairs, but also in a manner familiar within the linguistics literature, i.e. synthesis.

3.1 Meaning and Signal Spaces

The high-level model used a highly abstract representation of meaning and signal: a single integer value to represent one of three grammatical meanings, and the signal space was captured using two integers. One integer represented the original lexical origin of the form, and another represented its current phonetic form and usage conventions after repeated reanalyses were accrued.

Here, in the low-level model, both the meaning and signal spaces are modeled as collections of symbols. In the case of the signal space, these symbols are taken to represent individual phonemes, and in the meaning space, the symbols taken to represent sememes. There is no attempt to represent in more detail the phonetic or semantic features.

As in spoken language, phonemes occur one after the other, and so the signal space is stored as an ordered sequence, cf. a list or array in most modern programming languages. However, while semantic concepts do form hierarchies, they do not seem to depend on a particular order. To clarify, where “I ate the shark” and “The shark ate me” are very different signals, there is no such equivalent evidence that perceiving the color of the shark before the shape of the shark results in a different compositional structure. In particular, the semantic space is a multiset, which like a set has no notion of ordering, but does allow for items to be contained multiple times. For

example, consider the collections $\{a\ a\ b\ c\}$ and $\{c\ b\ a\ a\ a\}$. If these are taken as traditional sets, they would be identical, as after discounting order and repetition one would have $\{a\ b\ c\}$. However, as multisets, they would not be identical, as the second collection has 3 a's and the first only 2 a's.

The meaning and signal spaces, described above, are devised as an improvement on the those used in Kirby (2001), in which both meanings and signals were represented as flat sequences. For the signal space, a flat sequence makes sense: phonemes come in a linear sequence, and while participating in phonological hierarchies (e.g. syllable structure), in general they do not depend on them necessarily to convey meaning. For example, the sentence written here could be written out in IPA, and still be understood without hearing it pronounced.

In contrast, a flat sequence is a poor choice for a meaning structure. Meanings, as mentioned above, do not appear to be dependent on linear orders. Certainly processing of semantic information is affected by the order in which the elements of a scene are perceived, but once so processed, the relationship of percepts and objects is independent: looking at the gray fuzzy chair in my living room, attending to the grayness before the fuzziness does not alter my ability to identify or characterize the object. For this reason, an unordered structure is a more natural choice.

Further, a linear sequence of symbols does not allow for hierarchy to be represented. For example, when processing a room full of people at a party, each person has a location in the room, and each person in these locations has their own unique visual features. In order to process the scene, one needs to be able to keep these individual features associate with the people the

belong to, as well as where the people are in the room. In a linear sequence of sememes there is no way to group each applicable feature to just the person they belong to; if we list out the features in order, how are the groupings into people, and grouping of people into locations, represented. For this reason, in addition to an unordered structure, one that allows nesting into hierarchies is also a more natural choice.

Finally, just as an utterance can have multiple instances of the same phoneme, a meaning can have multiple instance of the same sememe: in the example of the room full of people, each location has many instances of the abstraction notion of objects, in this case human beings. A representation of meaning must allow for multiple instance to exist at the same hierarchical level of representation. If not, then when processing the room of people only one location of many in the room, and one person of many in that location would be perceivable.

Taking into consideration the above factors, it should be clear that a linear sequence, like that used to model a signal, is a poor choice for representing a meaning space. If we specify then a collection that is free of order, allows for hierarchy, as well as multiple instances of objects in the collection, the natural mathematical analog is the multiset; it meets all and only these criteria.

3.2 Grammar & Production

In order to model linguistic computation over the meaning and signal spaces described above, an abstract rewriting system (ARS) was chosen (Baader

& Nipkow, 1999). An ARS is specified over a collection of objects, and in addition to these, a collection of rules that enumerate the possible transformations of those objects. Each rule has a left-hand side and a right-hand side, where the structure in the left-hand side can be replaced, i.e. rewritten, with the structure on the right-hand side. In the low-level model, the objects transformed are meaning-signal pairings, and thus each rule specifies how a particular joint-pattern of meaning and signal symbols may be replaced with another. In order to rewrite, the meaning symbols specified in the left-hand side are removed from the multiset, and the right-hand side meaning symbols are added to it. Similarly, the contiguous sequence of signal symbols in the left-hand side are removed, and the right-hand side sequence is spliced in place.

In addition to the symbols that stand for the sememes and phonemes that constitute the linguistic content of the model, three more abstract kinds of symbols are used for computation. Each of the rewrite rule in the ARS has a single symbol from one of these three types as its left-hand side. This is similar to the restriction used in context-free grammars, where rewrite rules consist of a single non-terminal symbol on the left-hand side, and a sequence of non-terminal and terminals on the right-hand side. However, the ARS here is not strictly speaking a context-free grammar, as the objects of computation are not single symbol-sequences, but pairs comprised of a multiset and a sequence. In this respect, the ARS is more like that of a construction grammar, where linguistic content is built up from the repeated application of constructions, each specifying some joint semantic and signal transformation.

The first kind consists of a unique symbol, the **start symbol**, which only appears at the left-side of rewrite rules, and serves to indicate valid starting content for a derivation in the language. The second kind of symbols are the **morpheme symbols**, which are created during the first stage of the acquisition process (explained in the next section) and appear on the left-hand side of rules that rewrite to only sememe and phoneme content. These morpheme symbols represent the fundamental units of the meaning-signal space, and all rules with morpheme symbols as their left-hand side rewrite to content that cannot further be transformed. The final kind of symbols are the **transformation symbols**, created during the second stage of the acquisition process, and represent higher-level computational patterns in the language.

In order to use the ARS to produce a valid utterance, the agent must begin with a rewrite rule that has the start symbol as the left-hand side. Each such initial rewrite rule will have only transformation or morpheme symbols in the right-hand side. Additional transformations in the set of rules can be applied to these initials until either only sememe and phoneme symbols remain, which never appear in the left-hand side of a rule and thus no further computation is possible, or until the computational process itself is terminated externally. It is possible for recursive rules to be learned, as well as recursive chains of rules to develop, and so for practical purposes a limit is placed on the number of times a particular symbol may be rewritten, and the number of times each symbol is rewritten is monitored during production. If any symbol is rewritten a number of times equal to this limit, the derivation is aborted under the assumption that the system is stuck in an infinite loop.

Figure 3.1: Schematic of Compatible Signals

3.3 Language Acquisition

As input to learn the language, each agent receives a collection of meaning-signal pairs produced by other agents. This is analogous to the collection of signals that the agents in the high-level model used to determine their initial production set. Unlike in the high-level model, where production consisted of simply showing a signal to another agent, agents in the low-level model must induce an ARS from this raw input. This process consists of two stages, segmentation and pattern learning. In the first stage, the agents segment the language into best fundamental units of meaning and signal, and encode this segmentation in the ARS using rules headed by morpheme symbols. In the second stage, the agent considers the segmented grammar, searching for

computational patterns that best explain the segmentation, encoding these patterns in the ARS using rules headed by transformation symbols. At the end of this process, every initial input to the agent has been reduced to a sequence of morpheme and/or transformation symbols. At this point each such remaining unique initial input is encoded in the ARS as a rule headed by the start symbol. As a result of this two stage acquisition process, the agents are able to reproduce the input that was given to them without loss of coverage, and in addition may be able to generate novel productions not in the initial input, using generalizations from the pattern learning stage. The entire language acquisition process is summarized in Figure 3.1.

Before proceeding, the notion of ‘best’ as used in the statements “best fundamental units” and “best explain the segmentation”, of the input must be expanded on. Considering the segmentation of the input, there is no obviously correct way to divide the meaning and signal spaces. Moreover, even if one could determine if a given segmentation was optimal, there are too many possible segmentations to generate and test them all in a reasonable amount of time. Instead, the space of possible segmentation is searched incrementally using a heuristic-based approach to guide the process.

The heuristic taken here is the minimum description length, or MDL for short. The MDL approach was developed as the mathematical model of Occam’s Razor, which to paraphrase, is the idea that given a number of possible solutions, the simplest solution should be assumed. It has been applied to problems in a number of domains, broadly covered in Grünwald (2007). More than its breadth of applicability, MDL as a mathematical model for simplicity has been suggested as a potentially unifying principle

for all of cognitive science, underlying concepts in inference, vision, language, and many others (Chater, 1999; Chater & Vitányi, 2003). In particular, it has been demonstrated to account for open problems in language acquisition (?), as well as an ideal model for acquisition from positive evidence Chater & Vitányi (2007). This makes MDL based learning a clear starting point for modeling linguistic acquisition from usage examples like those in the low-level model.

To apply MDL to a problem, three things are required: a symbolic representation of the data, a means to compress that representation, and finally a function (usually called the heuristic or cost function) to numerically quantify a given representation and its compression. With these in place, one can search the space of possible compressions, working to minimize the cost. It may seem strange to talk about compression in relation to problem solving, but in general, a compression algorithm works by capturing recurrent patterns in a set of data. In this light, a more simple solution will consist of patterns that occur across large portions data, transforming individual instances into cases of more general patterns. As individual instances turn into cases of general patterns, they can be removed from the data, replaced by the single generalization they embody. This results in greater compression, and a more simple explanation of the data. As different generalizations and compressions of the data are generated, they can be compared quantitatively, and more complex solutions can be discarded in favor of more simple ones.

3.3.1 Segmentation

Returning to the problem of segmentation, we can define an MDL heuristic over our ARS used to capture an agent’s linguistic knowledge. We can start by transforming every meaning signal pair into a valid starting rule, namely the rule where the left-hand side is the start symbol, and the right-hand side the meaning-signal pair itself. By way of compression the ARS provides a natural choice, as each meaning-signal segment we identify can be captured as a rule where the left-hand side is a newly created morpheme symbol and the right-hand side the meaning and signal symbols we wish to segment. Then the data can be compressed by replacing the occurrence in the data of the segments meaning-signal symbols with the morpheme symbol of the rule; if we were to rewrite the data using our segmentation rule, we could recover the input exactly as before. The final piece needed for MDL is the cost function. A common approach is to develop a scheme to express the data and compression rules as a string of symbols, and then count the number of symbols in the string Grünwald (1995). For example, to represent the information content of a word, one could count the number of IPA symbols used to transcribe it.

With regard to the ARS rules, one such representation would be to take the components of a rule, and write them out directly with a dividing symbol after each component: the single abstract left-hand side symbol, the meaning sememe symbols, and the signal phoneme signals. With this representation, one can reconstruct the entire ARS from the string by reading until the first divider and assigning that content to the left-hand side, then reading until

the second divider and assigning that content to the right-hand side meaning, and finally reading until the third divider and assigning that content to the right-hand side signal. The process then begins over again to parse each additional rule until the entire ARS rule set had been recovered. With this in place, the three elements to use and MDL approach to evaluate possible solutions is in place. However, there remains the problem of how to search the space of possible solutions in a computationally efficient manner.

To search the space, a search procedure known as beam search is used (Russell & Norvig, 2016). To use a beam search, one needs two things: a way to incrementally step through the search space, and a way to compare steps against each other. The idea is that at any moment in time, the search is only considering a fixed number of possible solutions, which is called the beam width of the search. At each stage of the process, it increments each one of these candidate solutions each possible way, then evaluates these incremented states against each other, retaining only the top scoring candidates up to the limit set by the beam width. This process of incrementation, evaluation, and elimination continues until either the ideal candidate is found and the search concludes with an optimal solution, or the search runs out of possible ways to increment the candidates, and the search concludes with best solution it had discovered up to that point. This kind of search does not always lead to the discovery of the globally best solution, but allows for quickly finding good solutions in very large (potentially infinite) search spaces.

To search for segments, we will apply an MDL driven beam search. The process works by creating independent search frames that record the

meaning-signal symbols of a potential segment, and the places in the input where the segment occurs, which I will refer to as the support for the segment. Using the support of a search frame, we can increment the search frame by either possibly extending its meaning by one symbol, extending its signal by one segment, or determining that it can no longer be extended. In addition the MDL value of using the segment for compression can be calculated from just this local information, and so these search frames can be quickly compared to one another.

To begin, we calculate the cost of the input as encoded using the initial start rules described above. Then we will consider each symbol that occurs in the input and create a search frame for the three most common phonemes, and every position in the input where it occurs. To each of these search frames, every rule that contains the phoneme is added as support. Then, each unique sememe symbol from each meaning in the search frame’s support is considered. For each such meaning symbol, a new search frame is created with the already identified signal symbol, the newly identified meaning symbol, and the rules from the support where both the signal and meaning co-occur. At the end of this process, every possible way to create a minimal meaning-signal pair for the three most common phonemes is represented with a search frame.

To illustrate the search frame and its use in processing, consider the example in Figure 3.2. For simplicity, only the signal side of the search frame is shown above. In the model an analogous representation of the meaning space would also be present. On the top left we have the initial search frame, where the signal “cat” is under consideration so far, and the

Figure 3.2: Example of a Search Frame. For simplicity, only the signal side of the search frame is shown above. In the model an analogous representation of the meaning space would also be present.

attestations of it in the input are below in the support. Each support is split into the phonemes that occur before and after the considered signal. Here we can see that there are three possible ways to extend the signal based on the input: the phoneme symbols “y” (green) and “e” (blue) could be prefixed, or “o” (red) could be suffixed. Below, we can see the result of prefixing “y” or suffixing “o” (the prefixing of “e” is similar to result of “y” and has been omitted). In both cases, the signal under consideration has been updated to show the addition of the added phoneme. Likewise, the appropriate phoneme has been deleted from its location in the support.

To evaluate each search frame using MDL, we consider what the cost would be if the input were compressed with the meaning signal pair represented in the search frame. Specifically, the cost to create a new compression rule is equal to $4 +$ the number of symbols in the meaning and signal for the segment. The initial cost of 4 comes from the 3 dividing symbols used to encode the rule, and the 1 additional morpheme symbol created to label the left-hand side of the new rule. Since the support of the search frame consists of all the places in the input where the segment might occur, we know that each rule in the support would be rewritten. Since the meaning and signal sides would have the symbols of the new segment removed and replaced with the new morpheme signal used to represent the rule, the value of compressing the grammar using the segment is equal to the number of rules in the support, multiplied by the length of the segment’s signal and meaning minus two: $|rule| * (|meaning| + |signal| - 2)$. This means the total cost of using the compression to encode the grammar would be equal to the current cost of the input plus the cost of the compression rule, minus

the value gained by using the compression. At this stage each of the initial search frame candidates is evaluated, and the best (in the MDL sense) are retained up to the length of the beam width.

Next each search frame in the beam width is extended, if possible. As mentioned above, each search frame can either be extended with an additional meaning or additional signal symbol, and all such possible extensions can be quickly generated from the rules in the support. First, to generate extensions based on meaning, each unique meaning symbol in the meanings of the supports is considered. For each such symbol, a new search frame is generated with it now appended to the existing meaning and signal of the old search frame, and the support is updated to only retain rules where this new potential segment occurs. Similarly, to extend a search frame with a unit of signal, each possible symbol that is one beyond the existing signal of the search frame is considered, and a new frame generated with the signal thus extended and the support updated. Any search frames with no support are eliminated, as they would be unable to compress the input. Additionally, any extension that leaves a rule in the support without any input or any signal phonemes is removed from consideration, as if that segmentation were made there would be examples in the input where a morpheme is no longer possible.

Each extension is evaluated using the MDL heuristic, and compared with each other and the search frames already found in the previous iteration. If none of the extend search frames can minimize the cost beyond those already found, the search ends and the best candidate is selected. Otherwise, the extended search frames that are improvements are retained to the exclusion

of the previous candidates which they best, and the search continues for another iteration. Once the best candidate is selected, the input is compressed using the new segment.

After each compression, the remaining signal-meaning pairs in the input are checked against previously discovered segments to check if a segment discovered earlier has occurred elsewhere in the input. If so, the inputs are compressed and checked again, until no existing compressions can be applied. Then as before, the cost of the input is calculated, and all possible new initial segments are generated. The only difference being that after any morpheme symbols are ignored; the previous segments are not considered, only the remaining portions of the meaning-signal pairs that are currently unsegmented. The process continues until all the input had been segmented into morphemes.

3.3.2 Modified GRIDS: Extension of Idea

The learning in the model was an extension of the ideas presented in Langley & Stromsten (2000), in which a method of using MDL to learn context-free grammars was presented. The idea for the algorithm presented much earlier in (Wolff, 1982), but lacked a viable implementation. That is, conceptually the idea of alternating operations, one that abstract patterns and another that generalizes those abstractions was laid out in Wolff (1982), but the means of searching the space of abstractions/generalization, as well as deciding which among multiple abstractions/generalizations to make was a problem solved in Langley & Stromsten (2000). The so-called GRIDS

(GRAMMAR INDUCTION DRIVEN BY SIMPLICITY) system present by Langley et al. is the base from which the following modified version is developed.

As mentioned earlier, an ARS is not the same as context-free grammar; a context-free grammar is a special case of an ARS with certain restrictions. Of major limitation were the restrictions that a context-free grammar is defined over strings, there is only one space of inputs to search (i.e. just signal and no other), and that the inputs must already be segmented into fundamental units.

The solution to the first problem was mechanically presented in the previous section. If one wants to learn a rewrite rule system in which the most fundamental unit is a morpheme - defined in the model as any meaning-signal pair with at least one sememe and one phoneme - then the input needs to be segmented into this level of representation before the algorithm can begin. As previously described, to meet this challenge an MDL heuristic was developed, along with the search frames used to make the search computationally tractable.

A byproduct of the search procedure is that both the meaning and signal spaces have been recoded into an alternative alphabet; each abstraction is replaced with a new segmentation symbol (the symbols tagged SG in the low-level model appendix), and both the meaning and signal portions of the segmented input are replaced with it. While the structure of the meaning and signal spaces are mathematically different, multisets and sequences, respectively, they after segmentation they are now composed of the same fundamental symbols.

This leads to a solution to the problem of multiple spaces, as well as

the restriction to strings. The segmentation procedure, by definition, leaves exactly the same symbols on both the meaning and signal sides of an input. Thus, so long as any further operations continue to maintain this equality, the two spaces can effectively be treated as one. For example, if a sequence of symbols is ever abstracted from the signal side of an input, they same symbols must also be abstracted from the meaning. With this restriction that equivalence be maintained, either side can be operated on independently of the other, solving the multiple spaces issues.

A linear sequence of symbols is mathematically equivalent to a string, all be it that traditionally a string is made up of individual characters, one character per symbol. This makes the signal side of the ARS compatible with the ideas presented in the GRIDS system; the ability to make identical changes to either side of the inputs, means that GRIDS can be allowed to operate on the signal side in isolation, but whenever a generalization or abstraction is made, the appropriate adjustment to the meaning side is also codified in the ARS rule.

3.3.3 Modified GRIDS: Abstraction

After segmentation, the agents then evaluate the compressed input to discover meaningful patterns in the arrangement of the morphemes that can be further used to compress the ARS. The process here, as just described, is adapted from the GRIDS algorithm of Langley & Stromsten (2000), where it was used to create a context-free grammar from segmented string input, and successfully applied to learning English language syntactic patterns, as

well as the structures underlying chemical formula and genetic sequences. In general, the GRIDS algorithm works by alternating between two modes, abstract and generalize. In abstract, the algorithm searches for reoccurring patterns in the ARS, and uses this to compress the encoding further by creating new rules to capture them. In generalize, the algorithm searches for instances where abstract symbols (here the morpheme and transformation rules) may be merged. This has the effect of creating symbols that no longer stand for an instances of repeated information, but classes of information; the fact that the same symbol can be replaced multiple ways embodies the notion that each of these replacements functions in an interchangeable manner. This is equivalent in syntax to creating word classes, or in complex morphological systems, creating affix classes.

To abstract, the process begins by considering all rules that begin with a start symbol or transformation symbol; the fundamental meaning-signal pairs of sememe and phoneme symbols have already been dealt with during the segmentation process, and only higher level patterns are searched for here. Within these rules, every possible contiguous pair of symbols in the signal is generated, and a search frame begun with these two symbols, and the support where they occur. These search frames are evaluated using the same MDL heuristic for the segmentation search frames above, and the top scoring frames that fit within the beam width are retained. Then the input is scanned again for every possible contiguous triple of symbols, and corresponding search frames generated and evaluated using the MDL heuristic. If there are some triples that better compress the ARS than the pairs, they replace those search frames up to the beam width, and the search

is continued searching for contiguous quadruples. The process continues until such extensions cease to yield an improvement over existing search frames in the beam width.

Once the search completes, the top candidate is selected, and a new rule generated with a new transformation symbol as the left hand side, and meaning-signal symbols of the search frame as the right-hand side. After this the ARS is compressed using the new transformation rule, and a new search for abstractions begins. Eventually, no new abstractable pairs can be found, or the ones that can compress the ARS cost more to create than not. At this point the abstract operations concludes, and generalize is begun.

3.3.4 Modified GRIDS: Generalize

In order to generalize, i.e. merge, the abstract symbols of the ARS, two scenarios are considered. First each pair of ARS rules is searched to see if any two rules with identical right-hand sides can be found, but with different left-hand sides; that is, the ARS is searched to see if two identical meaning-signal patterns are being redundantly labeled. If so, the two left-hand side symbols are merged. In order to merge a pair of symbols, one of the left-hand sides of the two rules is chosen at random to remain in the ARS, and all instances of the other, wherever it occurs, is replaced by the symbol chosen to remain. The second way two symbols can be merged is if two symbols occur in the exact same environment, and thus represent a similar function. To do this, each pair of rules that begin with the same left-hand side are searched, and if their meaning and signals differ in a single position those

two different symbols are merged. This process continues until no more such generalizations can be found.

If a generalization was made, the abstract and generalize processes are run again. Often the merging of symbols in generalization leads to new patterns that can be abstracted. This alternation of abstraction and generalization continues until no more new rules can be formed. At this point the pattern extraction algorithm completes, and the resulting ARS is used by the agent to communicate with other agents.

In order to help visualize these procedures, some sample situations have been made and color-coded in Figure 3.3. The three ways the GRIDS operations can apply (abstract, and the two kinds of generalize) are shown top to bottom. On the left side of each operation is a mini set of ARS rules before the operation, and on the right side is the result of applying the operation to those rules. The coloring of symbols is consistent on both sides, so it is easier to see which symbols have been added or removed from the ARS rule set. In the ARS rules shown, only the signal component of the right-hand side is shown, because after segmentation, the meaning and signal components have the same symbols; the only difference is that order is meaningful in the signal component but does not matter for the meaning component.

On the top we have the abstract procedure, which is responsible for finding repeated patterns of information in the ARS and compressing them out by creating a new rule. Looking at the before rules (on the left), one such repeated pattern has been found using the beam search described above. In order to capture this pattern a new rule has been created, headed by the symbol R1 in red on the left-hand side. In the ARS all instances of the

S → SG1 SG4 SG5	S → SG1 R1
S → SG2 SG4 SG5	S → SG2 R1
S → SG3 SG4 SG6	S → SG3 SG4 SG6
S → SG4 SG5 SG7	S → R1 SG7
	R1 → SG4 SG5

abstract

S → SG3 R1	S → SG3 R2
S → SG2 R2	S → SG2 R2
R1 → SG4 SG5	R2 → SG4 SG5
R2 → SG4 SG5	

generalize content

S → SG3 R2	S → SG2 R2
S → SG2 R2	S → SG2 R2
R2 → SG4 SG5	R2 → SG4 SG5

generalize context

Figure 3.3: Abstract & Generalize

pattern have been replaced by this signal.

In the middle we have the generalize procedure, operating here on identical content (shown in blue) that has been classified using different symbols (in green and red). In this case the two different left-hand sides are singled out, and one of their symbols is chosen to be retained. In this example the symbol R2 (in red) is retained, and all instances of R1 (the symbol not retained, in green) have been replaced with R2. In addition the now identical compression rules have been merged.

Finally, on the bottom, we have the other instance when generalize can operate. Namely, when context suggests that two symbols are functionally equivalent. Here two different symbols, SG3 in blue and SG2 in purple, occur in the same context, in green. As with the other instance of generalize, one of the two selected symbols is retained, here SG2 in purple, and the rules are rewritten. Unlike with the previous example of generalization, the identical rules are retained as they are headed by the start symbol; they are a part of the data, and so are left in order that additional pattern finding with abstract may occur.

3.4 Processes of Morphological Development

3.4.1 Extension

In the high-level model, extension was implemented as a usage-based probability. Specifically, it was calculated on a signal per signal basis by multiplying the number of unique agents in the input that used the signal by

a small base probability. This small base probability represented the latent extension employed by a speaker for reasons such as expressiveness.

In the low-level model, extension is implemented based on attested examples of ‘analogy’ as used in the grammaticalization literature. The textbook example of such extension is the use of motion verbs to imply future time. In English, the phrase ‘going to’ was originally employed only literally to express travel to a destination for the purpose of an action. However, it began to be used in contexts where there was no such physical distance to traverse, only temporal, and over time it was semantically bleached to the point where it had no necessary physical connotation. This change is discussed in detail in Hopper & Traugott (2003).

In order to implement this behavior, periodically an agent will have a random sememe removed from the meaning of an utterance it produces. This creates exemplars in the input of learners where a signal does not have its literal meaning. Returning to the example of ‘going to’ we could imagine its original meaning represented as something like {away, travel, future, purpose}, and the extended meaning as something like {away, future, purpose} with the sense of ‘travel’ removed. The model does not include pragmatic concerns such as style or formality, and so the occurrence of this extension operation happens stochastically as determined with a model parameter.

3.4.2 Repetition Effects

As mentioned in Chapter 1 when discussing the creation of bound morphology, two kinds of repetition effects are particularly important. First, forms

that are highly frequent are targets for reanalysis as being dependent on or structurally part of grammatical classes in their surrounding context. Second, forms that are highly frequent are targets for reduction in phonetic form, as their identity is easily determined from context. The English ‘going to’ exemplifies both: it was phonetically reduced from ‘going to’ to ‘gonna’ over repeated shortenings, and in the ‘gonna’ form has been reanalyzed as functioning like an auxiliary verb due to its frequent occurrence in the same contexts such as ‘will’ and ‘shall’.

The first effect involving reanalysis is discussed in the next section proper. As for the second effect, the derivation tree of rewrites that produced each utterance an agent uses is saved, and a count of morphemes retained. For each utterance there is a small chance that the most used morpheme in the utterance will be reduced, and that chance is multiplied by the number of times the morpheme has been used at the time of the utterance. If a morpheme is selected for reduction, its signal representation is randomly elided by removing the first or last phoneme. An example of the basic phonological process of deletion. Once this reduction has occurred, the change to the signal representation is made permanently in the grammar of the agent, such that it will continue to use the reduced form in future productions.

3.4.3 Reanalysis

As used in the literature, reanalysis refers to a difference in the structural understanding of language patterns by a subsequent generation, such as (but not limited to) a change in the meaning a particular morpheme or grammat-

ical construction represents or the grammatical class of a morpheme. In the low-level model, reanalysis is a natural artifact of change in the input from one generation to the next. As discussed earlier in section 3.3, the agents segment the input, and then infer higher-order patterns using the MDL principle. Changes to the semantic structure happen during the segmentation phase.

Of particular importance for the development of bound morphology was the kind of semantic reanalysis, called ‘bleaching’ in the linguistics literature, whereby the meaning of a form becomes more general. That is, like with the ‘going to’ example, the meaning of the form is not just changed, but becomes less specific. This bleaching of a form’s meaning is often the result of an extension that has been reanalyzed to stand not for its literal meaning, but its extended one. In the low-level model, if such extensions occur in the input with enough frequency, it may become informatically more favorable to segment the the meaning-signal space using the reduced semantic content. That is, a few extended exemplars in the input might shift the MDL algorithm towards segmenting the reduced meaning-signal mapping rather than the literal interpretation of the prior generation.

The non-semantic connotation of reanalysis is also a natural result of the MDL algorithm. If a number of segments co-occur next to each other in the input, this may bias the MDL algorithm to create a single higher-order rewrite rule to capture this common pattern. It may be the case that despite such close co-occurrence of their signal representations, the derivations that produced these patterns in the grammar of the prior generation are quite different. As mentioned in the section above on repetition effects, ‘gonna’ has

come to be used like an auxiliary verb such as ‘will’ due to its occurrence in similar locations in the signal. However, derivationally speaking, ‘gonna’ is a contraction of a verbal phrase, inflected in the present progressive, which takes an infinitive object. An human child, or an agent in the model for that matter, would not have access to this derivation structure though when inferring the grammar that produced such utterances. In particular for the MDL algorithm, the simplest explanation of the meaning-signal space may be that ‘gonna’ is just another kind of modal verb, and even more so if this phonetically reduced form has been semantically bleached.

3.5 Measuring Synthesis

By definition, synthesis is the ratio of morphemes per word over a large representative corpus of a language. In order to calculate it then, one needs to be able to segment usage examples into morphemes, and further group those morphemes into words. While there is no standard definition for what constitutes a word in linguistics, proposals have been made using phonological information as the basis of the determination, or the grammatical structure as the basis for determination. Both are summarized in a definition proposed in Loos (2004):

A word is a unit which is a constituent at the phrase level and above. It is sometimes identifiable according to such criteria as:

- being the minimal possible unit in a reply
- having features such as:

- a regular stress pattern, and
 - phonological changes conditioned by or blocked at word boundaries
- being the largest unit resistant to insertion of new constituents within its boundaries, or
 - being the smallest constituent that can be moved within a sentence without making the sentence ungrammatical.

As the model here embodies no phonological content beyond the holistic use of symbols to represent phonemes, nor a model of discourse, the first two criteria cannot be used. However, the notion of resistance to insertion of new constituents, as well as the fact words are the topmost, movable components of an utterance, can be used to create a method of analysis. That is, given these constraints, and a representation in which morphemes are the smallest fundamental unit, this author proposes a novel method for classifying words and morphemes as presented below.

As a byproduct of the segmentation process, the morphemes of the language are entirely known for each agent. What remains to be done is to classify the morpheme and transformation symbols as representing syntactic transformations or word classes; once the word classes are identified, the number of words in an utterance can be quickly counted from the derivation tree, as well as the number of morphemes that are children of them.

The structural definition of a word yields a few constraints that can be used to examine the ARS of an agent. Beginning with the symbols used in

the top-level start rules, one can iteratively discover all the symbols which represent syntactic operations and those which represent word classes. The first structural restriction is that an utterance is composed of one or more words. For the ARS, this means that all symbols that can be rewritten to as the result of start rule must either stand for a word class, or stand for a syntactic class. We begin by assuming that all such symbols written by a start rule represent word classes, unless proven to violate an additional constraint that words must obey. If a symbol is shown to violate a constraint, it is then assumed to not be a word class, but a syntactic one (e.g. a phrase). As syntactic rules by definition decompose themselves into word classes and syntactic classes, we assume that all symbols generated by a syntactic class are word classes, until proven otherwise. This process continues until all symbols under consideration are shown to be word classes, and thus all the syntactic classes have been discovered. To clarify, once all the symbols standing for word classes have been discovered, then all symbols involved in rewriting the word classes must be syntactic in nature, and all symbols rewritten by the word classes must be morphological in nature.

The two constraints are drawn from the holistic nature of a word; a word is composed of one or more morphemes, but not other words or syntactic processes. The first constraint from this observation is that a symbol representing a word class cannot rewrite to itself, i.e. it is non-recursive. For example, a noun in the process of its derivation cannot split into two nouns; morphological transformations only add new morphemes into the existing, singular word's structure. The second constraint is that a symbol representing a word class cannot rewrite to the inside of another word, or outside of

itself into a new word. For example, in the process of producing a verb form, one of the verb's morphemes cannot be bound to another word, nor can one of the verb's morphemes separate into another word.

The procedure to check if a symbol violates one of these constraints is straightforward. First, check that each rule with the symbol as the left-hand side does not contain the symbol in the right-hand side, and thus is not recursive. Next check that for each rule where the symbol appears as the left-hand side, the symbols on the right-hand side are not rewritten to by a rule already identified as a word class or syntactic class. If the symbol does not violate one of these constraints, or if it only rewrites to symbols determined to be morphemes at segmentation, then it is a word class. By iteratively considering all symbols that are children of the start rules, all syntactic classes will be discovered, leaving the word classes that compose them clearly labeled, and and remaining symbols as sub-word morphemic classes and transformations. In addition once these classes have been discovered, the boundedness of words can be checked to see if any function as clitics. To illustrate the process, an ARS will be presented for two natural languages, and the classification of their rewrite rules using this algorithm worked in detail.

3.5.1 English & Lushootseed Examples

Using an ARS to capture language knowledge presents a number of analytic advantages for the study of synthesis. In particular, it allows for a straightforward calculation of synthesis using the ARS and the structural definition

for words, without recourse to a speaker of the language. This property is essential for simulation research, where by definition, there are no native speakers to be queried. To illustrate the method used, two simplified phrase structure grammars of natural languages will be used. The first language, English, provides a familiar and clear example of an analytic language (low degree of synthesis), and the second, Lushootseed, provides contrast with a polysynthetic language (high degree of synthesis) from native North America.

To begin with, let us present a simple subset of English syntax as a standard phrase structure grammar, a kind of rewrite rule system, seen in Figure 3.4. The following is not meant to be an exhaustive grammatical treatment of English, but merely one with a depth that keeps it compact and familiar, and a breadth that allows for illustration of the analytical method.

Rewrite rules headed by high-level symbols are presented at the top, and rewrite rules headed by symbols that capture morphemic content are presented at the bottom. Only the signal side of the rewrite rule is shown, as was explained above in the section on GRIDS, only one side of the meaning-signal pairs need be used.

In the above rules, the dollar sign is used to indicate the start symbol that all valid expressions generated by the grammar must begin with. The high-level symbols NP, VP, and AdjP stand for noun-phrase, verb-phrase, and adjective-phrase, respectively. The morphemic symbols, Det, Gen, N, V, and Adj, stand for determiner, genitive, noun, verb, and adjective, respectively. The genitive 's' is singled out as it is often analyzed as a clitic in literature on

$$\begin{aligned}
\$ &\longrightarrow \text{NP VP} \\
\$ &\longrightarrow \text{VP} \\
\text{NP} &\longrightarrow \text{N} \\
\text{NP} &\longrightarrow \text{Det NP} \\
\text{NP} &\longrightarrow \text{AdjP N} \\
\text{NP} &\longrightarrow \text{NP Gen N} \\
\text{AdjP} &\longrightarrow \text{AdjP Adj} \\
\text{AdjP} &\longrightarrow \text{Adj} \\
\text{VP} &\longrightarrow \text{V} \\
\text{VP} &\longrightarrow \text{V NP} \\
\\
\text{Det} &\longrightarrow \{\text{determiners ...}\} \\
\text{Gen} &\longrightarrow \{\text{s}\} \\
\text{N} &\longrightarrow \{\text{nouns ...}\} \\
\text{V} &\longrightarrow \{\text{verbs ...}\} \\
\text{Adj} &\longrightarrow \{\text{adjectives ...}\}
\end{aligned}$$

Figure 3.4: Simplified English ARS

grammaticalization, an intermediate step on the developmental cline from free morpheme to bound morpheme. This is not entirely standard within linguistics, as mentioned in the literature review, and the reader is referred to Klavans (2018) as an overview of the various proposals and current usages of the term clitic.

While the presentation in Figure 3.4 captures a typical phrase structure analysis of English, the ARS used in the model makes goes deeper than describing the computations for words, but also those for morpheme combinations below the word-level. In order to demonstrate how morphemic patterns can be included in a phrase-structure grammar, the rules above will be modified to specify the morphemic structure. The determiner and genitive rules are left unaltered, as these are traditionally analyzed as having holistic morphemic structure. In contrast the noun, verb, and adjective

$$\begin{aligned}
N &\longrightarrow RN \\
N &\longrightarrow RN \text{ IN} \\
V &\longrightarrow RV \\
V &\longrightarrow RV \text{ IV} \\
\text{Adj} &\longrightarrow R\text{Adj} \\
\text{Adj} &\longrightarrow R\text{Adj} \text{ IAdj} \\
\\
\text{Det} &\longrightarrow \{\text{determiners ...}\} \\
\text{Gen} &\longrightarrow \{\text{s}\} \\
\\
RN &\longrightarrow \{\text{noun-roots ...}\} \\
IN &\longrightarrow \{\text{s}\} \\
VR &\longrightarrow \{\text{verb-roots ...}\} \\
VI &\longrightarrow \{\text{s, ed}\} \\
R\text{Adj} &\longrightarrow \{\text{adjective-roots ...}\} \\
I\text{Adj} &\longrightarrow \{\text{er, est}\}
\end{aligned}$$

Figure 3.5: Morphological Revision

rules are refined as seen in Figure 3.5.

The convention above is to prefix R to a category to specify a root, and an I to specify an inflection. For example, an English noun may stand on its own or have some inflectional morphology added to it, such as the plural ‘-s’. The representation here supposes that after building the morpho-syntactic tree, further phonological processing would apply. For example, cat-s and dog-s would have identical RN IN representations, and at a later stage in production the plural suffix be differentiated into [s] and [z] realizations. Note, this is purely an example of a real-life phenomenon, and not one that exists in the model.

Recombining the syntactic and refined morphemic rules developed above, the simplified morpho-syntactic rewrite rule of English are expressed in Figure 3.6. The rules have been divided such that those that are syntactic

SYNTACTIC CLASSES

$\$ \longrightarrow \text{NP VP}$
 $\$ \longrightarrow \text{VP}$
 $\text{NP} \longrightarrow \text{N}$
 $\text{NP} \longrightarrow \text{Det NP}$
 $\text{NP} \longrightarrow \text{AdjP N}$
 $\text{NP} \longrightarrow \text{NP Gen N}$
 $\text{AdjP} \longrightarrow \text{AdjP Adj}$
 $\text{AdjP} \longrightarrow \text{Adj N}$
 $\text{VP} \longrightarrow \text{V}$
 $\text{VP} \longrightarrow \text{V NP}$

WORD CLASSES

$\text{N} \longrightarrow \text{RN}$
 $\text{N} \longrightarrow \text{RN IN}$
 $\text{V} \longrightarrow \text{RV}$
 $\text{V} \longrightarrow \text{RV IV}$
 $\text{Adj} \longrightarrow \text{RAdj}$
 $\text{Adj} \longrightarrow \text{RAdj IAdj}$

CLITIC CLASSES

$\text{Det} \longrightarrow \{\text{determiners ...}\}$
 $\text{Gen} \longrightarrow \{\text{s}\}$

MORPHEME CLASSES

$\text{RN} \longrightarrow \{\text{noun-roots ...}\}\text{s}$
 $\text{IN} \longrightarrow \{\text{s}\}$
 $\text{VR} \longrightarrow \{\text{verb-roots ...}\}$
 $\text{VI} \longrightarrow \{\text{s, ed}\}$
 $\text{RAdj} \longrightarrow \{\text{adjective-roots ...}\}$
 $\text{IAdj} \longrightarrow \{\text{er, est}\}$

Figure 3.6: Simplified English ARS (Including Morphological Revision)

transformations are grouped at the top, followed by those that are word classes next, then those that function as clitic classes¹, and finally those that are morphemic classes last.

To contrast against English, a language with a low degree of synthesis, Lushootseed, a language with a high degree of synthesis, will also be presented in the form of a simplified ARS. Lushootseed is a Salishan language, spoken around the Puget Sound in Washington State. As typical of many such polysynthetic languages, many utterances consist of a single, complex verb. Similar to the English example above, a simplified morpho-syntactic grammar is presented in Figure 3.7. As Lushootseed is not as widely known as English, some additional time will be spent explaining what the ARS symbols represent.

The details of Lushootseed verbal inflection are quite fascinating in their own right – it is often cited as an exception to the universal that all languages make a distinction between nouns and verbs (Beck, 2013). The basic syntactic structure of the language is presented in a two volume pedagogical grammar, Hess & Hilbert (1976), with morphological segmentation standardized in Bates et al. (1994) and Bierwert (1996). A detailed treatment of how these rules relate in Cognitive Grammar styled analysis, a kind of rewrite rule system, is developed further in Beck (1996), as is a corpus-based analysis of affix hierarchies presented in Beck (2009). The ARS presented in Figure 3.7 is based on these descriptive works.

The top group of rules are the syntactic transformations. The basic

¹In addition to these, Dixon (2007) proposes more than 60 additional items that can be considered intermediate between bound morphology and free lexical items in English.

$\$ \rightarrow V$
 $\$ \rightarrow V \text{ DP}$
 $\$ \rightarrow V \text{ AC}$
 $\$ \rightarrow V \text{ AC DP}$
 $\$ \rightarrow \text{DP}$
 $\text{DP} \rightarrow \text{D}$
 $\text{DP} \rightarrow \text{D V}$

 $\text{AC} \rightarrow \{\text{agent-clitics ...}\}$

 $\text{D} \rightarrow \text{Epi Dist}$
 $\text{V} \rightarrow \text{VPre VR VSuff}$
 $\text{V} \rightarrow \text{VR VSuff}$
 $\text{V} \rightarrow \text{VPre VR}$
 $\text{V} \rightarrow \text{VR}$

 $\text{VPre} \rightarrow \text{Gen Nom Asp}$
 $\text{VPre} \rightarrow \text{Gen Nom}$
 $\text{VPre} \rightarrow \text{Nom Asp}$
 $\text{VPre} \rightarrow \text{Asp}$
 $\text{VSuff} \rightarrow \text{LV S}$
 $\text{VSuff} \rightarrow \text{LV S VSuff}$

 $\text{VR} \rightarrow \{\text{verb-roots ...}\}$
 $\text{Gen} \rightarrow \{\text{possessives ...}\}$
 $\text{Nom} \rightarrow \{\text{nominalizers ...}\}$
 $\text{Asp} \rightarrow \{\text{aspectuals ...}\}$
 $\text{LV} \rightarrow \{\text{linking-vowels ...}\}$
 $\text{LS} \rightarrow \{\text{suffixes ...}\}$
 $\text{Epi} \rightarrow \{\text{epistemics ...}\}$
 $\text{Dist} \rightarrow \{\text{distals ...}\}$

Figure 3.7: Complete Simplified Lushootseed ARS

Lushootseed utterance minimally consists of a single, highly inflected verb (V), which has an implied patient or agent, depending on which verbal affixes are used. Optionally, the verb can be followed by a determiner phrase (DP) which overtly specifies the patient or agent inherent in the verb. The determiner phrase can be either a bare determiner (D), which details epistemic and distal information, or a determiner and additional inflected verb. This additional inflected verb typically has one of several possible affixes that serve to nominalize or reify the verb. Finally, in the first and second person, the verb can be followed by an agentive clitic (AC), which either replaces the determiner phrase in function, or can combine with the determiner phrase to change the focus of the utterance. This is listed in the second group.

The third group of rules are the word-classes, which consists of just the verb and determiner. All determiners are bi-morphemic, with the first morpheme specifying epistemic information (Epi), including visibility and definiteness, and the second morpheme specifying distance (Dist) from the speaker and hearer. Each verb consists of a root, which when used alone specifies either an implied patient in a state (without reference to the agent), or an implied agent carrying out an action (without reference to the patient, if any). That is, a bare root is either inherently unaccusative or unergative, as defined in Perlmutter (1978). In fact, the unaccusative/unergative semantics of verbal roots appears across the Salishan family, and has been proposed as a feature of Proto-Salishan (Davis & Matthewson, 2009). This bare verb root can take on a number of prefixes in fixed orders (VPre) and a potentially unlimited number of suffixes in a recursive process (VSuff).

The fourth group of rules do not have a clear label in English grammar,

and represent inflectional processes whereby the prefix and suffixes of the verb are arranged. The verbal prefix appears in the orderings listed, containing affixes specifying possession (Gen), nominalization and reification (Nom), and aspectual information (Asp). The suffixation process provides a range of details hard to capture in a single class. Every suffix consists of a linking vowel (LV) and suffix (S), which can be recursively added to the verb. Such suffix chains generally specify the shape or location of the patient when used with unaccusative roots and the direction or goal of the agent with unergative roots. Finally the individual morpheme classes are listed in the fifth and final group.

3.5.2 Worked Analysis

To analyze the English and Lushootseed ARS presented above, a novel, intermediate object is useful to expedite the process. We begin by transforming the ARS into a directed graph, here dubbed the **rewrite dependency graph**. The graph is constructed by adding a node for each symbol that heads the left-side of a rewrite rule, i.e. all abstract symbols. Then, proceeding rule by rule, a directed edge is added from the left-hand side symbol of the rule to each of the abstract symbols that appear on its right-hand side. Once complete, the graph then shows at a glance which rules can be rewritten to produce which symbols. It is important to be clear that an edge in the graph, say, from VP to N, does not imply that a VP always rewrites to an N. Such an edge in the graph simply notes that in the grammar there is at least one rule with VP on the left-hand side that has an N on the right-

hand side; it is a statement that an N can be potentially produced from the rewrite of a VP. Such rewrite dependency graphs have been produced from the English and Lushootseed ARS above, and can be seen in Figure 3.8 and Figure 3.9, respectively. We will use these to first analyze the English rule set, followed by the Lushootseed rule set.

The analysis process begins by starting from the rewrite dependencies of the start symbol, which has been placed physically at the top of each graph. We make the assumption that all children of the start symbol stand for either a syntactic transformation or a word class; an utterance consists of minimally a single word (regardless of the potential complexity of that word) or an arrangement of words. Similarly, by virtue of the segmentation process, all morphemes have already been identified by the symbols that rewrite exclusively to sememe and phoneme symbols. Each of these nodes at the bottom of the graph must either be a word or part of a word.

Once these nodes have been identified and labeled, we return to the children of the start symbol and work downward until all word classes have been identified and decided, and in the process, all of the syntactic transformations that arrange them. The procedure is to assume that all of the symbols encountered when working top-downward represent word classes, unless they violate one of the structural constraints of a word that forces them to be labeled as a syntactic transformation instead.

As explained above, the main constraints used in this determination are both based on the holistic nature of a word. In terms of the graph, a word class being holistic means that the node representing the word class roots the sub-tree of nodes below it. The first constraint is that no word class can

recursively refer to itself, for if it could potentially replicate itself, then it would not be holistic. For example, it would be impossible in the process of constructing a noun to add an affix that splits in from one noun into two. Similarly, the second constraint is that no child of a word class may be produced by a syntactic transformation or morphological process outside the sub-tree defined by the word class. Again, using the example of constructing a noun, it would be impossible to add an affix that is structurally bound to another word. Once all word classes have been identified, all symbols below a word in the graph are either morphemes, as identified in the segmentation process, or representative of transformations used in building words, e.g. the prefix orderings or recursive suffixation seen in Lushootseed.

The final step is to identify any clitics. Syntactically, a clitic functions as a word class, but is otherwise bound to another element in the utterance (Dixon, 2007). In terms of the rewrite rules, an abstract symbol is bound if its appearance is necessitated by the appearance of another symbol. Any word class that is bound is denoted as a clitic. As an example of such boundedness, looking at ARS for Lushootseed, we can see the possessive affixes (Gen) are bound to the nominalizing affixes (Nom), as a possessive can never appear without being followed by a nominalizer.

To begin the analysis of English, we identify the top-level nodes that are children of the start symbol:

Word Class or Syntactic Transformation { NP VP }

From this initial categorization, we begin to work downward until all word classes are identified. Here we assume that NP and VP are both

Figure 3.8: English Rewrite Dependency Graph

words unless they violate our constraints, and so must be taken as syntactic transformations.

Starting with VP, we see that it is not recursive, i.e. it cannot rewrite to itself, but that it does rewrite to nodes outside of its sub-tree. Specifically, it rewrites to NP which is also a rewrite of the start symbol, and it additionally rewrites to Adj, which is a rewrite of AdjP. As this violates the constraint that a word class must root the sub-tree below it, VP cannot be a word and must be a transformation. Moving on to NP, we see that it is recursive, and thus also cannot be a word class.

Now that we have determined that both NP and VP are syntactic transformations, their children must also be either word classes or additional syntactic transformations. We update our classification of nodes and pro-

ceed:

Syntactic Transformation { NP VP }

Word Class or Syntactic Transformation { Gen Det N AdjP V }

We see that Gen and Det are bottom-level symbols in the graph, i.e. they only rewrite to sememe and phoneme content. We now know their parent, NP, is a syntactic transformation, and must rewrite to either a word class or another syntactic transformation. However, both Gen and Det cannot possibly rewrite to any abstract symbols capable of representing words or additional transformations; they must stand for word classes. In contrast, we see that AdjP is both recursive, and rewrites to N, which is also a possible rewrite from NP; AdjP must be a syntactic transformation and not a word class. Finally, neither N nor V violate any constraints, and so we continue the assumption that they are words.

Syntactic Transformation { NP VP AdjP }

Word Class { Gen Det N V }

Word Class or Syntactic Transformation { Adj }

The only remaining child of a syntactic transformation is Adj. It is not recursive, and is the root of its sub-tree, and therefore we maintain the assumption that it is a word class and not a phase. Now, all the top-level nodes in the graph have been assigned either as a word class or syntactic transformation, which means the remaining nodes either reflect individual

morphemes or transformations that occur in the process of constructing words themselves.

The remaining task is to check if any word class symbols are bound, and should thus be classified as clitics instead. Looking at the rewrite rules, we see that both Gen and Det always occur with NP and only in this context. They are bound to NP, and so we label them as clitics, giving us the final classification that matches the intuitive categorization given above. It should be reiterated that these rewrite rules were processed solely on their rewrite relationships, and not on any inherent background knowledge of English; any other symbols with such a structure would lead to the same analysis

Syntactic Transformation { NP VP AdjP }

Word Class { N V Adj }

Clitic { Gen Det }

We can conduct the same process for the Lushootseed example as well, and a rewrite dependency graph is presented in Figure 3.9. To begin, we identify the top-level nodes that are children of Root.

Word Class or Syntactic Transformation { V DP AC }

Starting with V, we see that it is a valid root of its sub-tree, and is not recursive; V is a word class. AC as a child of the start symbol, must be either a word class or syntactic transformation. However, like Det and Gen in the English example, it only rewrites to morphemic content, and so the only option is that it too is a word class. Finally, while DP is not recursive,

Figure 3.9: Lushootseed Rewrite Dependency Graph

it fails to root its sub-tree by writing to V outside itself, and thus DP must be a syntactic transformation.

Syntactic Transformation { DP }

Word Class { AC V }

Word Class or Syntactic Transformation { D }

Continuing, we already know that one of the two children of DP, V, is a word, which leaves only the status of D undetermined. It violates no word constraints, and so is assumed to be a word. Finally, we check the word classes for potential clitics. Looking at the rewrite rules, we see that AC is

bound to V, and therefore a clitic. The analysis is complete:

Syntactic Transformation { DP }

Word Class { V D }

Clitic { AC }

3.6 Simulations

In the previous sections a low-level model of language structure and grammatical knowledge was presented. The meaning-signal pairs of integer values from the high-level model were replaced with multisets and sequences of symbols, respectively. Instead of using flat probabilities for the occurrence of extension, reanalysis, and repetition, these processes were implemented directly on the structure of the meaning-signal pairs themselves. Instead of choosing signals based on frequency alone, the use of the minimum description length principle was employed as a proxy for human learning, and an abstract rewrite system was inferred from the raw input received by an agent. Finally, instead of counting the sustained occurrences of processes that support bound morphological development, a method measuring the level of synthesis directly was proposed. Using the structural properties of a word, an automated procedure to determine which rewrite rules represent word classes and morphemes was used to calculate the synthesis of language in a network.

The simulations in this chapter follow closely Simulation 3 and Simulation 4 of the high-level model, which, respectively, investigated the effect of the clustering coefficient in isolation, and then the effect of social orga-

nization as seen in ecologically valid models of small and large societies. While the effect of size was studied independently in Simulation 5, it will be incorporated directly herein.

Methods

The same random, 25-agent networks from Simulation 3 of the high-level model, and corresponding 125-agent networks from Simulation 5 were used. Following the convention of Kirby (2001), the meaning space consisted of 25 possible structures, each made up of two sememes. The sememes were divided into two groups, A and B, and each group had five associated symbols, i.e. $\{A1\ A2\ A3\ A4\ A5\}$ and $\{B1\ B2\ B3\ B4\ B5\}$. Each meaning had one symbol from the A group and one symbol from the B group, e.g. $\{A2\ B5\}$. The signal space consisted of 25 phoneme symbols, comparable the number of phonemes in a natural language like Hawaiian, labeled P1, P2, and so on through P25. A signal could be 2 to 8 phonemes in length.

A singular initial language was created in which each of the 25 possible stimuli was assigned a randomly generated signal, and the ARS consisted of one start rule: $S \rightarrow R1\ (R1)$. Then for each of the randomly generated stimulus-signal pairs, a rule $R1 \rightarrow \text{stimulus (signal)}$ was added. Thus like in the simulations of the high-level model, each agent began with linguistic knowledge void of bound morphology; each rule is an instance of a single-word class R1, which rewrites directly to the meaning-signal content.

The construction of each agent also closely mirrors that of the high-level model. Each agent consisted of a history of meaning-signal pairs successfully communicated to them, and active and passive repertoires of meaning-

signal pairs used for communication and understanding. Unlike the high-level model, each agent also possess an ARS that represents their language knowledge.

Language change in the low-level model was also divided into diffusion and transmission stages. Each diffusion consisted of 10 rounds of communication, in which each agent in randomized order would select an interlocutor from their neighbors in the graph. To communicate, the agent would randomly select a meaning from the 25 possible stimuli, and produce a meaning-signal pair from their active repertoire. Then their interlocutor would search their passive repertoire for a mutually intelligible signal with the same meaning. Two meaning-signal pairs were deemed mutually intelligible if the meanings matched, and the signals were different by at most one signal; i.e. signals with an edit-distance of one were accepted. If no match could be found, the speaker could try to repair the communication using another signal with the same meaning from their active repertoire.

If a compatible meaning-signal pair could be found either in passive knowledge or through active repair, the hearer added the meaning-signal pair to their history, and additionally to their passive repertoire if not already present. Additionally, there was a small chance (same as in the high level model, $p = 0.25$) that this successfully communicated signal would be also added to the hearer’s active repertoire. Conversely, if a communication could not be established, a new random signal was generated for the meaning, and this new meaning-signal pair was added to both the active and passive repertoires of the speaker and hearer. Once this exchange is complete, the hearer and speaker switch roles, with the new speaker selecting

a random meaning to communicate; each randomly selected communication results in both agents receiving new input.

Intergenerational transfer, the transmission stage, is where the new direct implementation of extension, repetition, and reanalysis come to bear. Each agent induces a new ARS from the meaning-signal input they received during the diffusion stage. First the language is segmented as described above, and then the GRIDS algorithm is allowed to run to completion to extract meaningful patterns and generalizations.

Once the ARS is induced, the agent uses it to produce a new signal for each meaning of the 25 possible stimuli. The derivation tree is retained, and there is a chance ($p = 0.1$) that one of the meaning signal pairs will be operated on by extension (implementing the mechanics of semantic bleaching as described above). Then for each morphemic symbol there is a small probability ($p = 0.001$), multiplied by the number of times it appears in the derivation tree, that it will undergo erosion. The 25 meaning-signal pairs produced are then updated to reflect any semantic or phonemic changes that may have occurred due to extension and repetition. If these processes cause any meaning-signal pair to lose all of its signal, it is replaced with a new random signal.

At this point in the transmission stage, each agent has induced a new ARS from the raw input and produced meaning-signal pair for each stimulus. The algorithm described above is then run over the derivation trees produced by the agents, and the number of morphemes and words are calculated for each stimulus. For each agent in the network the synthesis score is taken as defined in the linguistics literature, $synthesis = \frac{\#morphemesincorpus}{\#wordsyncorpus}$ and

then averaged over all agents in the network. This gives a direct measure of the synthesis of the language used by agents in the network at generational transfer.

Before moving further, a brief discussion of the parameter values should be made. With regards to the diffusion of language across the network, the parameters were set as close as possible to those from the high-level model, so as to increase the comparability of the results; the changes to the structure of the language in the low-level model most directly effect the transmission process, i.e. how the effects of extension, repetition, and reanalysis are handled, and so the diffusion procedure was kept as much as possible. The only significant change in parameter settings between diffusion in the high-level model and the low-level model is with respect to the concept of distance between meaning signal pairs. In the high-level model the integer value for reanalysis was allowed to vary by ± 1 between meaning-signal pairs. In the low-level model, a difference in meaning or signal space is not expressed as an integer, but by the constituent phoneme/sememe features. To keep the distance measure as close as possible, meaning-signal pairs in the low-level model are allowed to be ± 1 sememe or phoneme symbol and still be considered mutually intelligible. Similarly, the probability a meaning may be bleached was kept at the same value in the high-level model for extension effects to occur.

While not simulated yet, we can reason about how changing these parameters may affect the results of the model. First, if we allow the distance between meaning-signal pairs to be greater than ± 1 feature, agents will accept more radically changed meaning-signal pairs as expressing the desired

meaning. This should reduce the effect of hub agents or the bottlenecks of intergenerational transfer. If an innovation can be more radical and still understood, then it is more likely to be communicated as an alternative across hubs rather than being replaced by a new, paraphrastic construction. Similarly, as discussed in the parameter exploration section of Simulation 5 in the high-level model, increasing the probability for extension effects should increase the number of variant meaning-signal pairs in the network. Every time an extension is made, the meaning or signal side of a pair loses one or more features. These extended forms, when entering the input, provide opportunities for reanalysis in which the semantically bleached or phonologically eroded version for a given meaning is learned as fundamental.

3.6.1 Simulation 6: Low-Level Transitivity

As mentioned just above, the same random networks from Simulations 3 and 5 from the high-level model were reused. Each agent was initialized with the same start grammar in all replications. Each simulation was allowed to run for 200 generations, with 10 rounds of communication across the network between each generation. Each network topology underwent 96 replications. The average synthesis was recorded after each transmission, and shown in Figure 3.10 and Figure 3.11.

A mixed-effects model was fitted with individual networks as a random effect, and the global clustering coefficient, as a fixed effect, to predict average levels of synthesis over the 10 network topologies. The mean level of synthesis was selected from the last 100 generations (generations 100

Figure 3.10: Results of Simulation 6, 25-Agent Networks: On the x-axis is the generation number as measure of time, and the average synthesis of the language on the y-axis. Shaded regions are the LOESS estimates supplied by the ggplot2 package in R.

Figure 3.11: Results of Simulation 6, 125-Agent Networks: On the x-axis is the generation number as measure of time, and the average synthesis of the language on the y-axis. Shaded regions are the LOESS estimates supplied by the ggplot2 package in R.

through 200), as only the stable behavior of the models at the end of their evolutionary path was of interest. A model was fit for each network size.

For the 25-agent networks, Figure 3.10, the results were similar to those obtained in the high-level version, Simulation 3: the statistical model yielded a positive effect of the clustering coefficient ($\beta = 0.06$; $t=73.69$; $p<.001$), with more dense networks able to sustain languages with higher levels of synthesis. Also similar, there was nonlinear effect of increased density on synthesis; linear increases in the clustering coefficient values of the networks does not result in equal increases in synthesis values.

For the 125-agent networks, Figure 3.11, the results were also similar to those obtained in Simulation 5: the statistical model again yielded a positive effect of the clustering coefficient ($\beta = 0.01$; $t=16.78$; $p<.001$), with more dense networks able to sustain languages with higher levels of synthesis. For greater details of the statistical results, please see the R reports appendix.

Also like in the high-level model, in both network sizes, a number of networks cluster tightly together at similarly high values: networks with clustering greater than 0.3 in the 25-agent networks, and greater than 0.1 in the 125-agent networks. In addition, those that do not cluster tightly together appear to stabilize at lower levels of synthesis. However, unlike with Simulation 5, only the network with connection probability 0.1 is unable eventually reach this common level of synthesis.

It should be noted that for this small meaning space the maximal value of synthesis is theoretically 2.0, as there only two kinds of sememe features (those labeled A and B) as described above in the model section. Since the meaning side of a morpheme is minimally a single sememe, there can

be at most 2 morphemes used to represent a meaning in the model. Thus, if each meaning in the language was represented by a word class in the model, and each word had the maximal 2 morphemes used to compose it, the over all synthesis value would be 2.0. After the initial fluctuation of synthesis values, the networks stabilize, on average across the runs, at the given levels. Within each network there is a mixture of agents who have obtained the maximum value for synthesis, but others who have not, even though they are expressing the same mutually intelligible surface forms. This is similar to the high-level model in which agents may possess some meaning-signal pairs with high reanalysis values along with others with much lower values, and bringing down the average.

3.6.2 Simulation 7: Low-Level Social Structure

We have just the results from the comparable high-level simulations concerning the effect of the clustering coefficient, similarly, we will now examine the networks from Simulations 4 and 5 from the high-level model concerning the effects of social structure. In addition, the random network with clustering coefficient of 0.4 was included, as this also the clustering coefficient for Hierarchical. Each agent was initialized with the same start grammar in all replications, and the same start grammar from Simulation 6 was used again. Each simulation was allowed to run for 200 generations, with 10 rounds of communication across the network between each generation. Similar to the high-level model, the 125-agent networks showed much less variability between runs, and so only 64 replications were conducted. The average syn-

Figure 3.12: Results of Simulation 7, 25-Agent Networks: On the x-axis is the generation number as measure of time, and the average synthesis of the language on the y-axis. Shaded regions are the LOESS estimates supplied by the ggplot2 package in R.

Figure 3.13: Results of Simulation 7, 125-Agent Networks: On the x-axis is the generation number as measure of time, and the average synthesis of the language on the y-axis. Shaded regions are the LOESS estimates supplied by the ggplot2 package in R.

thesis was recorded after each transmission, and presented in Figure 3.12 and Figure 3.13.

As in Simulation 6, a mixed-effects model was fitted with individual networks as a random effect, and the global clustering coefficient, as a fixed effect, to predict average levels of synthesis over the 3 network topologies. Data was selected from the last 100 generations (generations 100 through 200), as only the stable behavior of the models at the end of their evolutionary path was of interest. A model was fit for each network size. Additionally a second mixed-effects model was fitted with individual networks as random effect, topology and network size (25 versus 125 agents), and their interactions as fixed effects.

For the 25-agent networks, Figure 3.12, again the results were similar to those obtained in the high-level version, Simulation 4: The model yielded main effects of topology, such that the Complete yielded larger levels of synthesis. In particular, Complete developed significantly higher reanalysis than Hierarchical ($\beta = 0.02$; $t=16.10$; $p<.001$), as did the random network with identical levels of clustering ($\beta = 0.01$; $t=5.13$; $p<.001$).

For the 125-agent networks, 3.13, the results were also similar to those obtained in Simulation 5: The model yielded main effects of topology, such that the Complete again yielded larger levels of synthesis ($\beta = 0.02$; $t=31.50$; $p<.001$), as did the random network with identical levels of clustering ($\beta = 0.02$; $t=27.11$; $p<.001$). Interestingly, the synthesis levels of Hierarchical appear to be unaffected by the increase of size.

This may be explained better with the results of the second model that combined both sizes to investigate the interaction between size and topology.

The model yielded main effect of topology, such that Complete and the 0.4 networks yielded significantly larger levels of reanalysis ($\beta = 0.02$; $t=18.65$; $p<.001$). Unlike with Simulation 5, though an effect of size on reanalysis was not obtained ($\beta = -0.001$; $t=-1.3$; $p=0.19$). However, the topology:size interaction was significant ($\beta = 0.01$; $t=2.575$; $p<.01$), and is shown in Figure 3.14. For greater details of the statistical results, please see the R reports appedix.

The nature of this pattern, as well as the results from Simulation 6 suggest that as in the low-level model the effect of size is to increase the effect of the global clustering coefficient on synthesis. However, Hierarchical is the only network that does not support greater synthesis with the increase in size, rather it remains the same. This is reminiscent of results from Simulation 5 in which the increase in size of Hierarchical in the high-level model did create an initial improvement in the ability to support repeated reanalysis. However, this initial boost was eventually undone by the leveling action of hubs in the topology. Again, the results in the low-level mode suggest that additional layers of hierarchy in the social network serve to limit synthesis of a language and its speakers.

3.7 Discussion & Conclusion

The most important aspect of the low-level model results are not the statistical significance: in a simulation, if one possessed the resources to increase the sample size and number of replications, a p-value can be easily tuned. The effect size is also difficult to evaluate, as the meaning space (both in

Figure 3.14: Topology:Size Interaction: The mean levels of reanalysis from generations 100 to 200 are shown with 95% confidence intervals.

terms of size, and physical structure) is not on par with what one would expect in an accurate model of human language.

However, what is important, then, is the fact that the qualitative pattern of results mirrors that of the high-level model. This demonstrates the role the global clustering coefficient plays in supporting the processes required for bound morphology do actually predict the outcome when measured in a model tracking bound morphology itself. To my knowledge, no other model has tried to probe the development of bound morphology by breaking it down into its sub-process, let alone embed them in a social network. Even so, the high-level model only makes predictions about the necessary conditions for bound morphology to emerge. The low-level model takes this one step

further, seeing if these necessary conditions actually lead to higher levels of synthesis.

Again there is no reason the qualitative performance of the various networks should be similar to those of the high-level model. In the high-level model, there were three key differences from the low-level model. First, the language representation does not allow for the direct measure of synthesis. Second, the process that supports bound morphology received their values directly from the network structure — some might describe them as “baked in” to the model. Third, the model tracks the number of times these processes occur in tandem, as necessary for bound morphology to develop, but this is not sufficient to make claim about the presence of such bound morphology.

The low-level model addresses these concerns directly. It provides a compositional signal and meaning space, over which grammars can be learned and analyzed to calculate morphemes, words, and their proportion. In addition, the three processes that support the development of bound morphology are completely unrelated to the network structure for their function. Finally, the model measures synthesis directly; while the high-level model tracked the potential support for developing bound morphology, the low-level model measured if that predicted support indeed is actualized in practice. With this in mind, let us then summarize the qualitative performance of the high-level and low-level networks.

In the random networks of Simulation 6, the effect of increased clustering led to a similar, non-linear increase of synthesis, as predicted by Simulation 3. Similarly, in Simulation 7 the synthesis levels of the hierarchical network

lagged behind both the complete network and random network with an identical clustering coefficient. This too was predicted by the reanalysis levels of Simulation 4. Finally in line with the results of Simulation 5, increasing the size of the networks produced a similar amplification of the effect of clustering. And, again, Hierarchical was resistant to this expected increase.

One difference though is pronounced effect of the increasing the size from 25 agents to 125 agents in the random networks. In the high-level model, the gap between topologies that were able to perform at the highest levels and those that stabilized at lower levels appeared to widen. However, in the all low-level model, all 125-agent random networks seemed to increase in performance as compared to their 25-agent counterparts. Still, the 0.1 clustering network stabilizes at a lower value, but the rest seem to function similarly. This may be due to the small meaning space used, and perhaps with a larger range of potential synthesis values, differences would appear.

The low-level model also suggests that the translation of the three processes responsible for bound morphology to a low-level of representation is a viable basis for future work. In particular, the ability of MDL to capture segmentation and reanalysis may suggest the nature of more domain-general learning processes at work. Even so, the level of detail in the low-level model pales in comparison to that typically discussed in traditional linguistic description, and begs future expansion.

One such potential avenue for immediate investigation concerns the meaning space used by the stimuli. Unlike with natural language, there was no composition present; each stimulus consisted of one group of sememes,

whereas in most utterances there are multiple elements that would each be represented. In this sense, the meaning space here is most like an object naming game, in which each meaning describes the features of the object (Kirby et al., 2008). Similarly, the holistic phonemes of the signal space could be modeled as collections of features, and more subtle phenomena such as morpho-phonology could be included. This would be of particular interest to historical linguists working on sound change and comparative method.

While not from the model itself, another such potential expansion is the algorithm used to measure the synthesis levels of the language automatically from the ARS. While the examples from English and Lushootseed were provided to give a sense of how the algorithm operates, there is no reason that a proper corpus of segmented natural language could not be used to validate it. For many languages, such as English, segmentation and syntactic analysis tools exist that could be employed.

Chapter 4

Conclusion

In this thesis, I have attempted to unify a number of research strands, all with connections to the study of how social network structure affects the typological distribution of bound morphology. First, the idea from Trudgill (2011) that intimacy of speakers may be the underlying social force in language change has been operationalized quantitatively. This was done by taking the early Sociological work done in Rapoport (1977) linking the intimacy of speakers to apparent changes in the social network. The main result was that triangular relationships form between intimates, and so it was reasoned that in a network where there is in general more intimacy there should be more such triangular relationships. There already exists a quantitative measure for this property of a social network, the global clustering coefficient, and so this thesis proposed that as a numerical proxy for the much more vague notion of human intimacy.

Second, while prior computational models such as Brighton et al. (2005) and Kirby (2001) have been conducted to examine how compositionality may

emerge in the course of human evolution, they are incapable at examining the linguistic details directly. For example, ? used a vector space model, and while elegantly developing a measure to estimate relative magnitudes of compositionality, was unable to distinguish morphemes and words, and by extension, unable to measure synthesis itself. Similarly, the model of ?, while having discrete meaning and signal spaces, failed to capture fundamental structural difference between meaning and signal representations.

It should be mentioned, that these failings are not necessarily negligent on the part of the modelers mentioned, but that their goals in modeling are not linguistic in nature. They are modeling language, for sure, but at a much more abstract level, specifically with the goal of exploring how humans acquired the language faculty itself. One such major goal of this research is to understand what conditions are necessary for compositionality to emerge. In contrast, the level of detail in the two models of this work were taken from the perspective of a linguistic, and designed with a purpose to answer questions related to linguistic theories of language change. The work in this thesis shows how such computational approaches can be adapted to traditionally non-computational areas of linguistic theory.

Third, this thesis makes clear the need to differentiate the kinds of changes that affect the accumulation of innovation horizontally across populations, versus the kinds of changes that accumulate vertically through intergenerational transfer. In particular, the interaction between horizontal changes that affect the input used by learners during acquisition have a direct impact on the structural changes that lead to bound morphology. Prior work, such as that in Reali et al. (2014) and expanded in Ralli (2012) conflate

the incremental changes that lead to bound morphology with the one-shot changes that happen in contact between speakers of the same generation: in their model acquiring new forms using bound morphology happens in the same manner as acquiring new lexical forms, probabilistically through contact, where the difference is forms utilizing bound morphology have a lower probability to spread.

While there model is certainly valid in showing how novel forms may spread from one adult to another in a social network, this kind of grammar borrowing is not how bound morphology develops in the first place. While an excellent model of how as social networks grow, forms with complex morphology may be less favored, it speaks nothing about how such complex forms develop in the first place. This explanation of how languages may become more grammatically simple over time is common in the literature, cf. the Linguistic Niche Hypothesis proposed in Lupyan & Dale (2010) and developed in Lupyan & Dale (2016).

The work in this thesis begins to address, computationally, the issue of how social network structure affects to development of bound morphology in the first place. It examines the linguistics literature on how bound morphology has been observed to develop in natural language, and identifies three key cognitive processes – extension, repetition, and reanalysis - necessary for its formation. It then takes a unique approach, and rather than trying to model bound morphology directly, models the processes themselves at two levels of detail: the high-level and low-level models.

The low-level model examines how the clustering coefficient of networks modulates the capacity of the network to support these processes. As a

result, it was discovered that all things equal, the clustering coefficient, as a proxy for intimacy, predicts the support for bound morphology to develop. Further, it was discovered, that when looking at network topologies that embody the peculiar characteristic of large human social networks, namely, that degree of an agent is inversely proportional to the clustering around the agent, that the effect of intimacy seems to be undone by the emergence of hub agents. As hubs appear in larger social networks, they restrict the flow between smaller, more densely connected areas of the network where complexity can easily develop. However, the results from the high-level model are only speculative with regard to synthesis. The high-level model did not capture synthesis itself, just the support of the necessary conditions for it to develop. Moreover, the operationalization of the key cognitive processes was tied to the network structure itself. To overcome these limitations, the low-level model was developed.

It provides a compositional signal and meaning space, over which grammars can be learned and analyzed to calculate morphemes, words, and their proportion. In addition, the three processes that support the development of bound morphology are completely unrelated to the network structure for their function. Finally, the low-level model measures synthesis directly; while the high-level model tracked the potential support for developing bound morphology, the low-level model measured if that predicted support indeed is actualized.

It must be cautioned, thought, that this work is very much in initial stages, and certainly ignores many important aspects of natural language structure. For instance, even in the low-level simulation all of phonology

and its related processes are abstracted over with a symbolic representation. Also, there is no model of social dynamics to incorporate other well-studied aspects of social networks, such as prestige or identity, which are also known to affect language development. However, that said, these limitations can certainly be overcome with additional refinements in future studies. Indeed, one advantage the low-level model has over those such as Brighton et al. (2005) and Reali et al. (2018) is the nestable, feature-based structure of its architecture. Many linguistic theories are specified in terms of collections of features, and thus the same fundamental model structure may be more easily adapted to investigate a number of phenomena.

Also, as was mentioned at the end of Chapter 4, the kind of simulation work done in the thesis is not meant to provide an analytical model of human systems. That is, the goal of such agent-based approaches is to discover how the collection of simple mechanisms that underly a system produce the complex, emergent effects when combined together. Such simulation work is designed to explore the mechanisms not so the simulation can be analyzed, but so that they may lead to traditional analytical or experimental work where no such clear path existed before. In this regard, the results of this thesis contribute several avenues for further investigation.

For example, the result that hubs seem to restrict the flow of information leads to a potential avenue for analytical investigation. Mathematically, one could view the connections of the social network as communication channels, in the sense of information theory, and the role clustering plays as representative of noise in the system. A certain amount of such noise (caused by extension, repetition, and reanalysis) is required for morphological complex-

ity to develop, but too much leads to instability. Taking the information theory approach, one could try and convert the probabilistic modeling of those processes in the high-level model into a prediction for the amount of such noise each node of the network may contribute, and from there work out the channel capacity of the edges coming into each node as well. This would allow for a much more compact representation of the model: instead of as agent-based model, as a Markov chain Monte-Carlo process. This kind of numerical model is easily captured efficiently on GPU units, and could lead to the exploration of much larger, perhaps even realistic sized, social networks. Similarly, one may be able to analyze the output of the model analytically.

Turning away from mathematic extensions, the notion of the hub as restricting morphological complexity and high clustering as responsible for supporting it, human experiments can be devised. Iterated learning has been a recent mainstay in language evolution work (Kirby et al., 2008), but is difficult to execute with even population sizes of 20 - 30 subjects, as the size of the network determines how many subjects are required for a single generation of learning, and tens of generations are required to observe effects (Smith et al., 2003). However, the simulation results suggest that it is the distribution of clustering driving the effects on bound morphology, and much smaller network sizes (minimally 5 agents as show in (Ravasz & Barabási, 2003)) could be used to test these features specifically.

Bibliography

- Baader, F. & Nipkow, T. (1999). *Term rewriting and all that*. Cambridge university press.
- Bakker, D. (2010). *The Oxford handbook of linguistic typology*. Oxford Univ. Press.
- Barabási, A.-L. (2014). Network science book. *Network Science*.
- Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439), 509–512.
- Bates, D., Hess, T., & Hilbert, V. (1994). *Lushootseed dictionary*. University of Washington Press.
- Bauer, L. (2003). Introducing linguistic morphology.
- Baxter, G. J., Blythe, R. A., Croft, W., & McKane, A. J. (2006). Utterance selection model of language change. *Physical Review E*, 73(4), 046118.
- Beck, D. (1996). Transitivity and causation in lushootseed morphology. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 41(2), 109–140.

- Beck, D. (2009). A taxonomy and typology of lushootseed valency-increasing suffixes. *International Journal of American Linguistics*, 75(4), 533–569.
- Beck, D. (2013). Unidirectional flexibility and the noun-verb distinction in lushootseed. *Access mode: <http://www.ualberta.ca/~dbeck/FlexDist.pdf>*.
- Bierwert, C. (1996). *Lushootseed texts: an introduction to Puget Salish narrative aesthetics*. U of Nebraska Press.
- Brighton, H. & Kirby, S. (2001). The survival of the smallest: Stability conditions for the cultural evolution of compositional language. In *European Conference on Artificial Life*, (pp. 592–601). Springer.
- Brighton, H., Smith, K., & Kirby, S. (2005). Language as an evolutionary system. *Physics of Life Reviews*, 2(3), 177–226.
- Bromham, L., Hua, X., Fitzpatrick, T. G., & Greenhill, S. J. (2015). Rate of language evolution is affected by population size. *Proceedings of the National Academy of Sciences*, 112(7), 2097–2102.
- Campbell, L. (2000). What’s wrong with grammaticalization? *Language sciences*, 23(2-3), 113–161.
- Campbell, L. (2013). *Historical linguistics*. Edinburgh University Press.
- Carneiro, R. L. (1967). On the relationship between size of population and complexity of social organization. *Southwestern Journal of Anthropology*, 23(3), 234–243.

- Chakrabarti, M., Heath, L., & Ramakrishnan, N. (2017). New methods to generate massive synthetic networks. *arXiv preprint arXiv:1705.08473*.
- Chambers, J. K. & Schilling, N. (2013). *The handbook of language variation and change*, volume 129. John Wiley & Sons.
- Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *The Quarterly Journal of Experimental Psychology: Section A*, 52(2), 273–302.
- Chater, N. & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in cognitive sciences*, 7(1), 19–22.
- Chater, N. & Vitányi, P. (2007). ‘ideal learning’ of natural language: Positive results about learning from positive evidence. *Journal of Mathematical psychology*, 51(3), 135–163.
- Comrie, B. (1989). *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.
- Cook, S. F. & Heizer, R. F. (1968). Relationships among houses, settlement areas, and population in aboriginal california. *Settlement archaeology*, 79–116.
- Croft, W. (2002). *Typology and universals*. Cambridge University Press.
- Davis, H. & Matthewson, L. (2009). Issues in salish syntax and semantics. *Language and Linguistics Compass*, 3(4), 1097–1166.

- De Boer, B., Sandler, W., & Kirby, S. (2012). New perspectives on duality of patterning: Introduction to the special issue. *Language and cognition*, 4(4), 251.
- Dixon, R. M. (2007). Clitics in english. *English Studies*, 88(5), 574–600.
- Dryer, M. S. & Haspelmath, M. (Eds.). (2013). *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.
- Dunbar, R. (1998). The social brain hypothesis. *Evolutionary Anthropology*, 6(5), 178–190.
- Erdős, P. & Rényi, A. (1959). On random graphs, i. *Publicationes Mathematicae (Debrecen)*, 6, 290–297.
- Evans, N. & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, 32(05), 429–448.
- Fischer, O., Norde, M., & Perridon, H. (2004). *Up and down the cline—the nature of grammaticalization*, volume 59. John Benjamins Publishing.
- Fortescue, M. D. (1984). *West Greenlandic*. Croom Helm London.
- Gabelentz, G. v. d. (1894). Hypologie der sprachen, eine neue aufgabe der linguistik. *Indogermanische Forschungen*, 4, 1.
- Giannakis, G. (1993). The tense-aspect system of the indo-european verb. *Word*, 44(3), 485–495.
- Givón, T. (1979). *On Understanding Grammar*. New York: Academic Press.

- Gonçalves, B., Perra, N., & Vespignani, A. (2011). Modeling users' activity on twitter networks: Validation of dunbar's number. *PloS One*, 6(8), e22656.
- Granovetter, M. S. (1977). The strength of weak ties. In *Social networks* (pp. 347–367). Elsevier.
- Greenberg, J. H. (1960). A quantitative approach to the morphological typology of language. *International journal of American linguistics*, 26(3), 178–194.
- Grünwald, P. (1995). A minimum description length approach to grammar inference. In *International Joint Conference on Artificial Intelligence*, (pp. 203–216). Springer.
- Grünwald, P. D. (2007). *The minimum description length principle*. MIT press.
- Han, C.-h., Musolino, J., & Lidz, J. (2016). Endogenous sources of variation in language acquisition. *Proceedings of the National Academy of Sciences*, 113(4), 942–947.
- Haspelmath, M. (1997). *From space to time: Temporal adverbials in the world's languages*. Lincom Europa.
- Haspelmath, M. (2001). *Indefinite pronouns*. OUP Oxford.
- Haspelmath, M. & Sims, A. (2013). *Understanding morphology*. Routledge.
- Heine, B. (1993). *Auxiliaries: Cognitive forces and grammaticalization*. Oxford University Press.

- Heine, B. (1997). Possession: Cognitive sources. *Forces, and Grammaticalization, Cambridge.*
- Heine, B. (2003). On degrammaticalization. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, 163–180.
- Heine, B. & Kuteva, T. (2002). *World lexicon of grammaticalization*. Cambridge University Press.
- Herrera, C. & Zufria, P. J. (2011). Generating scale-free networks with adjustable clustering coefficient via random walks. In *Network Science Workshop (NSW), 2011 IEEE*, (pp. 167–172). IEEE.
- Hess, T. M. & Hilbert, V. (1976). Lushootseed: An introduction, books 1 and 2. *Seattle, WA: American Indian Studies, University of Washington.*
- Hickey, R. (2010). *The handbook of language contact*. John Wiley & Sons.
- Himmelmann, N. P. (2004). Lexicalization and grammaticization: Opposite or orthogonal. *What makes grammaticalization*, 21–42.
- Hockett, C. F. & Hockett, C. D. (1960). The origin of speech. *Scientific American*, 203(3), 88–97.
- Hopper, P. J. & Traugott, E. C. (2003). *Grammaticalization*. Cambridge University Press.
- Jiang, N. (2004). Morphological insensitivity in second language processing. *Applied Psycholinguistics*, 25(4), 603–634.

- Jin, E. M., Girvan, M., & Newman, M. E. (2001). Structure of growing social networks. *Physical review E*, 64(4), 046132.
- Ke, J., Gong, T., & Wang, W. S. (2008). Language change and social networks. *Communications in Computational Physics*, 3(4), 935–949.
- Ke, J. & Yao, Y. (2008). Analysing language development from a network approach. *Journal of Quantitative Linguistics*, 15(1), 70–99.
- Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2), 102–110.
- Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31), 10681–10686.
- Klavans, J. L. (2018). *On clitics and cliticization: The interaction of morphology, phonology, and syntax*. Routledge.
- Labov, W. (2007). Transmission and diffusion. *Language*, 83(2), 344–387.
- Langley, P. & Stromsten, S. (2000). Learning context-free grammars with a simplicity bias. In *European Conference on Machine Learning*, (pp. 220–228). Springer.
- Loos, E. E. (2004). *Glossary of linguistic terms*. SIL International.
- Lupyan, G. & Dale, R. (2010). Language structure is partly determined by social structure. *PloS One*, 5(1), e8559.

- Lupyan, G. & Dale, R. (2016). Why are there different languages? the role of adaptation in linguistic diversity. *Trends in cognitive sciences*, 20(9), 649–660.
- Luraghi, S. & Bubenik, V. (2010). *Continuum companion to historical linguistics*. Bloomsbury Publishing.
- Martin, J. H. & Jurafsky, D. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall.
- McCarty, C., Killworth, P. D., Bernard, H. R., Johnsen, E. C., & Shelley, G. A. (2001). Comparing two methods for estimating network size. *Human organization*, 60(1), 28–39.
- Meillet, A. Linguistique historique et linguistique générale.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Milroy, J. & Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Journal of linguistics*, 21(2), 339–384.
- Milroy, L. (1987). Language and social networks.
- Milroy, L. & Milroy, J. (1992). Social network and social class: Toward an integrated sociolinguistic model. *Language in society*, 21(1), 1–26.
- Narrog, H. & Heine, B. (2011). *The Oxford handbook of grammaticalization*. Oxford University Press.

- Nettle, D. (1999). Using social impact theory to simulate language change. *Lingua*, 108(2-3), 95–117.
- Nettle, D. (2012). Social scale and structural complexity in human languages. *Phil. Trans. R. Soc. B*, 367(1597), 1829–1836.
- Newman, M. (2010). *Networks: an introduction*. Oxford university press.
- Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2), 167–256.
- Newmeyer, F. J. (2000). Deconstructing grammaticalization. *Language sciences*, 23(2-3), 187–229.
- Niyogi, P. & Berwick, R. C. (1997). A dynamical systems model for language change. *Complex Systems*, 11(3), 161–204.
- Norde, M. (2009). *Degrammaticalization*. Oxford University Press.
- Nowak, M. A., Krakauer, D. C., & Dress, A. (1999). An error limit for the evolution of language. *Proceedings of the Royal Society of London B: Biological Sciences*, 266(1433), 2131–2136.
- Perlmutter, D. M. (1978). Impersonal passives and the unaccusative hypothesis. In *annual meeting of the Berkeley Linguistics Society*, volume 4, (pp. 157–190).
- Ralli, A. (2012). Morphology in language contact: verbal loanblend formation in asia minor greek (aivaliot). *Morphologies in contact (Studia Typologica)*. Berlin: Academie Verlag, 185–201.

- Ramat, A. G. & Hopper, P. J. (1998). *The limits of grammaticalization*, volume 37. John Benjamins Publishing.
- Rapoport, A. (1977). Contribution to the theory of random and biased nets. In *Social Networks* (pp. 389–409). Elsevier.
- Ravasz, E. & Barabási, A.-L. (2003). Hierarchical organization in complex networks. *Physical Review E*, 67(2), 026112.
- Reali, F., Chater, N., & Christiansen, M. H. (2014). The paradox of linguistic complexity and community size. In *Evolution of Language: Proceedings of the 10th International Conference (EVOlang10)*, (pp. 270–277). World Scientific.
- Reali, F., Chater, N., & Christiansen, M. H. (2018). Simpler grammar, larger vocabulary: how population size affects language. *Proc. R. Soc. B*, 285(1871), 20172586.
- Refsing, K. (1986). *The Ainu language: the morphology and syntax of the Shizunai dialect*. Aarhus Univ. Press.
- Revere, P. & William, P. (1994). *The evolution of grammar: Tense, aspect and modality in the languages of the world*. Chicago: University of Chicago Press.
- Russell, S. J. & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.
- Sapir, E. (1921). *Language: An Introduction to the Study of Speech*. New York: Harcourt Brace & Company.

- Schleicher, A. (1859). *Zur Morphologie der Sprache*, volume 1. K. Akademie der wissenschaften.
- Sebba, M. (1997). *Contact languages: Pidgins and creoles*. Macmillan.
- Seifart, F. (2012). The principle of morphosyntactic subsystem integrity in language contact: Evidence from morphological borrowing in resígaro (arawakan). *Diachronica*, 29(4), 471–504.
- Senghas, A. & Coppola, M. (2001). Children creating language: How nicaraguan sign language acquired a spatial grammar. *Psychological science*, 12(4), 323–328.
- Smith, K., Brighton, H., & Kirby, S. (2003). Complex systems in language evolution: the cultural emergence of compositional structure. *Advances in Complex Systems*, 6(04), 537–558.
- Smith, K., Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2008). Introduction. cultural transmission and the evolution of human behaviour.
- Smith, K. & Kirby, S. (2008). Cultural evolution: implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1509), 3591–3603.
- Soffer, S. N. & Vazquez, A. (2005). Network clustering coefficient without degree-correlation biases. *Physical Review E*, 71(5), 057101.
- Svorou, S. (1993). *The Grammar of Space, Amsterdam/Philadelphia*. PA: John Benjamins. Google Scholar.

- Trudgill, P. (2002). *Sociolinguistic Variation and Change*. Georgetown University Press.
- Trudgill, P. (2011). *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford University Press.
- Varga, I. (2015). Scale-free network topologies with clustering similar to online social networks. In *Proceedings of the International Conference on Social Modeling and Simulation, plus Econophysics Colloquium 2014*, (pp. 323–333). Springer.
- Von Humboldt, W. (1823). *Ueber das Entstehen der grammatischen Formen, und ihren Einfluss auf die Ideenentwicklung*. Druckerei der Königl. Akademie der Wissenschaften.
- von Schlegel, A. W. (1818). *Observations sur la langue et la littérature provençales, par AW de Schlegel*. A la librairie grecque-latine-allemande, rue des Fosses-Montmartre.
- von Schlegel, F. (1808). *Über die Sprache und Weisheit der Indier*. Mohr und Zimmer.
- Wodak, R., Johnstone, B., & Kerswill, P. E. (2010). *The SAGE handbook of sociolinguistics*. Sage Publications.
- Wolff, J. G. (1982). *Language acquisition, data compression and generalization*. Pergamon.
- Zuidema, W. & De Boer, B. (2009). The evolution of combinatorial phonology. *Journal of Phonetics*, 37(2), 125–144.

Chapter 5

Appendix: Diachronic

Low-Level Example

To better illustrate the typical evolution of grammatical knowledge three snapshots from a single agent will be presented. The snapshots were taken at generation 1, generation 42, and generation 69, at which point the agent first obtained synthesis levels of 1.0, 1.5, and 2.0, respectively. Each snapshot contains the ARS grammar rules induced by the agent, as well as the rewrite dependency graph of these rules.

In order to keep the ARS grammars and accompanying graphs small enough to be easily understood, the meaning space of the agents was reduced: in the low-level simulations used in the thesis, the sememes used in the meaning space were divided into two groups, A and B, and each group had five associated symbols, i.e. $\{A1\ A2\ A3\ A4\ A5\}$ and $\{B1\ B2\ B3\ B4\ B5\}$. Each meaning had one symbol from the A group and one symbol from the B group, e.g. $\{A2\ B5\}$. In what follows, the meaning space is still divided

into two groups, A and B, but each group has only two symbols.

The network structure in which the agent was embedded was a random network, constructed with 25 agents at a connection probability of 0.1. The simulation was run for 100 generations of transfer, with 5 rounds of communication per generation.

Generation 1

At generation 1, the agent has seen one round of communication. As seen in Figure 5.1 the agents grammar consists of 4 top-level rules, two of which (1 and 3) that rewrite to whole words, and two of which (2 and 4) that use syntactic means to combine individual words to express their meanings. Each meaning in the grammar is expressed as either an indivisible word or combination of two indivisible words, and so naturally the synthesis level of any corpus using this grammar is 1.0.

	symbol	meaning	signal	word
1:	S	SG3	SG3	no
2:	S	SG2-SG4	SG2-SG4	no
3:	S	SG1	SG1	no
4:	S	SG2-SG5	SG2-SG5	no
5:	SG1	a2-b1	p6	yes
6:	SG2	b2	p11	yes
7:	SG3	a1-b1	p6-p3	yes
8:	SG4	a1	p4	yes
9:	SG5	a2	p5-p8	yes

Figure 5.1: Grammar at Generation 1. (Top) Rewrite Dependency Graph
(Bottom) ARS Rules

Generation 42

As seen in Figure 5.2, by generation 42 the agents grammar has increased in complexity, acquiring some bound morphology, and achieving a synthesis level of 1.5 overall. The grammar consists of 3 top-level rules, 2 rules (1 and 3) rewrite to individual words, and 1 rule (2) that rewrites to a complex word form.

	symbol	meaning	signal	word
1:	S	SG5075	SG5075	no
2:	S	R116	R116	no
3:	S	SG5078	SG5078	no
4:	SG5075	a1-b2	p7-p2-p2-p11-p9	yes
5:	SG5076	b2	p5	no
6:	SG5077	a1	p2	no
7:	SG5078	a2-b1	p4	yes
8:	SG5077	a2	p10	no
9:	R116 SG5076-SG5077	SG5076-SG5077	SG5076-SG5077	yes

Figure 5.2: Grammar at Generation 42. (Top) Rewrite Dependency Graph
(Bottom) ARS Rules

Generation 69

Finally by generation 69, as seen in Figure 5.3 the agents grammar has reached the maximum level of synthesis possible of 2.0. The grammar consists of 2 top-level rules, both of which rewrite to complex word forms. These two word classes are distinguished by the prefix in the signal that rewrites to either sememe a1 or a2. Each word class then has a suffix that expresses both b1 and b2 differentially.

	symbol	meaning	signal	word
1:	S	R231	R231	no
2:	S	R230	R230	no
3:	SG8769	a2	p11-p8	no
4:	SG8770	a1	p1	no
5:	SG8771	b1	p7	no
6:	SG8771	b2	p9-p10-p8	no
7:	SG8773	b1	p3-p9	no
8:	SG8773	b2	p12-p4	no
9:	R230	SG8769-SG8773	SG8769-SG8773	yes
10:	R231	SG8770-SG8771	SG8770-SG8771	yes

Figure 5.3: Grammar at Generation 69. (Top) Rewrite Dependency Graph
(Bottom) ARS Rules

Chapter 6

Appendix: R Reports and Model Details

In the following appendix the R outputs for the various LMEM models is provided for reference.

In terms of software, the analysis was conducted in the R Studio development environment, running R version 3.5.1. Plotting, with the exception of network data, was handled using the `ggplot2` package, and data manipulation was done both in vanilla R for smaller data sets, and the `data.table` package for data sets of more than 5 GB in size. The network measures were analyzed using the `igraph` package, and network plots were also generated using this package. The LMEM models themselves were done using the `lme4` package, along with the `lmerTest` package to add supplemental information to the vanilla `lme4` model outputs.

Chapter 2

2.3

Linear mixed model fit by REML. t-tests use Satterthwaite's method [`'lmerModLmerTest'`]

Formula: `complexity ~ topology + (1 | network_id_unique)`

Data: `transitivity_average[transitivity_average$generation >= 1000, ]`

REML criterion at convergence: 6610.9

Scaled residuals:

Min	1Q	Median	3Q	Max
-4.5389	-0.2538	-0.0005	0.2396	4.4791

Random effects:

Groups	Name	Variance	Std.Dev.
network_id_unique	(Intercept)	9.958	3.1557
Residual		0.127	0.3563

Number of obs: 2000, groups: network_id_unique, 1000

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	7.0511	0.2163	997.9996	32.605	<2e-16 ***
topology	3.2646	0.3485	997.9996	9.367	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

(Intr)

topology -0.886

2.4

Linear mixed model fit by REML. t-tests use Satterthwaite's method [`'lmerModLmerTest'`]

Formula: `complexity ~ topology + (1 | network_id)`

Data: `social_dt[social_dt$generation == 1000]`

REML criterion at convergence: 46890.8

Scaled residuals:

Min	1Q	Median	3Q	Max
-4.0002	-0.6276	-0.0370	0.5052	3.8179

Random effects:

Groups	Name	Variance	Std.Dev.
network_id	(Intercept)	1.208	1.099
Residual		6.172	2.484

Number of obs: 10000, groups: network_id, 100

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	5.83413	0.12061	129.95782	48.373	< 2e-16 ***
topologycomplete	4.26960	0.07027	9897.00001	60.762	< 2e-16 ***
topologyrandom	-0.21693	0.07027	9897.00001	-3.087	0.00203 **
topologybarabasi	-5.08880	0.07027	9897.00001	-72.421	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

	(Intr)	tplgyc	tplgyr
toplgycmplt	-0.291		
topolgyrndm	-0.291	0.500	
topolgybrbs	-0.291	0.500	0.500

2.5

25-Agent and 125-Agent Social Networks

Linear mixed model fit by REML. t-tests use Satterthwaite's method [`'lmerModLmerTest'`]

Formula: `complexity ~ topology * size + (1 | network_id)`

Data: `df[generation == 500, ]`

REML criterion at convergence: 1400

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.6359	-0.4592	-0.0263	0.4468	2.8966

Random effects:

Groups	Name	Variance	Std.Dev.
network_id	(Intercept)	0.6143	0.7838
Residual		5.7094	2.3894

Number of obs: 300, groups: network_id, 100

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	6.5299	0.2515	287.2950	25.967	< 2e-16 ***
topologycomplete	3.6463	0.3379	181.0920	10.790	< 2e-16 ***
size125	-5.4106	0.4199	225.1718	-12.885	< 2e-16 ***
topologycomplete:size125	4.5149	0.5853	181.0920	7.714	7.88e-13 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

	(Intr)	tplgyc	siz125
toplgcmplt	-0.672		
size125	-0.541	0.402	
tplgcm:125	0.388	-0.577	-0.697

125-Agent Random Networks

Formula: complexity ~ topology + (1 | network_id_unique)

Data: transitivity_averages[transitivity_averages\$generation >= 500,]

REML criterion at convergence: 2744.1

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.75072	-0.26242	-0.03738	0.26463	2.97460

Random effects:

Groups	Name	Variance	Std.Dev.
network_id_unique	(Intercept)	3.3198	1.8220
Residual		0.1221	0.3495

Number of obs: 1000, groups: network_id_unique, 500

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	2.5901	0.1776	498.0001	14.58	<2e-16 ***
topology	8.7972	0.2863	498.0001	30.73	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

(Intr)

topology -0.886

3.6.1

25-Agent Networks

Linear mixed model fit by REML. t-tests use Satterthwaite's method [`'lmerModLmerTest'`]

Formula: `synthesis ~ topology + (1 | replication)`

Data: `df1[size == 25 & generation >= 100, ]`

REML criterion at convergence: -294077

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.7535	-0.6941	-0.0225	0.6707	4.2366

Random effects:

Groups	Name	Variance	Std.Dev.
replication	(Intercept)	1.668e-05	0.004085
	Residual	6.011e-03	0.077532

Number of obs: 129280, groups: replication, 128

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	1.378e+00	5.894e-04	4.895e+02	2337.95	<2e-16 ***

topology 5.532e-02 7.507e-04 1.292e+05 73.69 <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

(Intr)

topology -0.701

125-Agent Networks

Linear mixed model fit by REML. t-tests use Satterthwaite's method [`'lmerModLmerTest'`]

Formula: `synthesis ~ topology + (1 | replication)`

Data: `df1[size == 125 & generation >= 100, ]`

REML criterion at convergence: -252598.4

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.9838	-0.6693	-0.0040	0.6679	4.1933

Random effects:

Groups	Name	Variance	Std.Dev.
replication	(Intercept)	7.077e-07	0.0008412
	Residual	1.175e-03	0.0342774

Number of obs: 64640, groups: replication, 64

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	1.418e+00	3.096e-04	6.743e+02	4580.35	<2e-16 ***
topology	7.875e-03	4.694e-04	6.458e+04	16.78	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

	(Intr)
topology	-0.834

3.6.2

25-Agent Networks

Linear mixed model fit by REML. t-tests use Satterthwaite's method [`'lmerModLmerTest'`]

Formula: `synthesis ~ topology + (1 | replication)`

Data: `df2[df2$size == 25 & df2$generation > 100, ]`

REML criterion at convergence: -87194.2

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.6001	-0.6999	-0.0221	0.6707	4.3297

Random effects:

Groups	Name	Variance	Std.Dev.
replication	(Intercept)	1.743e-05	0.004175
Residual		6.027e-03	0.077632

Number of obs: 38400, groups: replication, 128

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	1.407e+00	7.791e-04	5.426e+02	1805.783	< 2e-16 ***
topology1.0	1.562e-02	9.704e-04	3.827e+04	16.097	< 2e-16 ***
topology0.4	4.980e-03	9.704e-04	3.827e+04	5.132	2.88e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

	(Intr) tp11.0
topology1.0	-0.623
topology0.4	-0.623 0.500

125-Agent Networks

Linear mixed model fit by REML. t-tests use Satterthwaite's method [`'lmerModLmerTest'`]

Formula: `synthesis ~ topology + (1 | replication)`

Data: `df2[df2$size == 125 & df2$generation > 100, ]`

REML criterion at convergence: -74465.9

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.9152	-0.6681	-0.0068	0.6636	4.0656

Random effects:

Groups	Name	Variance	Std.Dev.
replication	(Intercept)	4.471e-18	2.114e-09
Residual		1.209e-03	3.477e-02

Number of obs: 19200, groups: replication, 64

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	1.406e+00	4.346e-04	1.920e+04	3235.63	<2e-16 ***
topology1.0	1.936e-02	6.146e-04	1.920e+04	31.50	<2e-16 ***
topology0.4	1.666e-02	6.146e-04	1.920e+04	27.11	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

	(Intr)	tpl1.0
topology1.0	-0.707	
topology0.4	-0.707	0.500

Combined 25-Agent and 125-Agent Networks

Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']

Formula: synthesis ~ topology * size + (1 | replication)

Data: df3[generation > 100,]

REML criterion at convergence: -98486.8

Scaled residuals:

Min	1Q	Median	3Q	Max
-4.1340	-0.6005	-0.0104	0.5734	4.9830

Random effects:

Groups	Name	Variance	Std.Dev.
replication	(Intercept)	9.892e-06	0.003145
	Residual	4.492e-03	0.067021

Number of obs: 38400, groups: replication, 128

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	1.407e+00	6.544e-04	4.757e+02	2149.967	<2e-16 ***
topology1.0	1.562e-02	8.378e-04	3.824e+04	18.646	<2e-16 ***
size125	-1.375e-03	1.052e-03	1.066e+04	-1.308	0.191
topology1.0:size125	3.736e-03	1.451e-03	3.824e+04	2.575	0.010 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

	(Intr)	tpl1.0	siz125
topology1.0	-0.640		
size125	-0.510	0.398	
tplg1.0:125	0.370	-0.577	-0.690

Chapter 7

Appendix: Pseudo-Code

Chapter 2

2.1.1

Communicate Round (network)

 randomize network's connections

 FOR EACH connection

 diffuse the two agents

 END

END

Diffuse (agent_a, agent_b)

 IF one agent has innovated and the other has not

 IF chance ≤ 0.1

 set both agents to have innovated

```
        END
    END
END
```

2.1.2

```
Transmission ( network )
    changes := store for updated innovation status
    FOR EACH agent IN network
        add updated status to changes
    END

    FOR EACH each change and index in changes
        set the corresponding agent at that index to change
    END
END
```

```
Update Status ( agent )
    IF chance <= number of agent's connections that have innovated
        RETURN innovate
    END
    RETURN not innovated
END
```

```
Communicate Round ( network )
```

```

    randomize network's connections
  FOR EACH connection
    diffuse the two agents
  END
END

Diffuse (agent_a, agent_b)
  IF one agent has innovated and the other has not
    IF chance  $\leq$  0.1
      set both agents to have innovated
    END
  END
END

```

2.3 - 2.5

```

Transmission ( network )
  FOR EACH agent IN network
    replace actives with result of intergenerational transger
    replaces passives with new actives
    remove all old experiences from history
  END
END

```

```

Intergenerational Tansger ( agent )

  chosen := store for selected signals

  FOR EACH meaning in language
    select the signal used by the most unique agents for the meaning
    in case of ties, choose amongst them at random
    add signal to chose
  END

  FOR EACH signal in chosen
    IF chance <= 0.01
      replace the realization with the next id number
      increase reanalysis by 1
    END
  END

  FOR EACH signal in chose
    IF reanalysis level >= 20
      replace signal origin, and realization with next ids for each
      set reanalysis level to 0
    END
  END

  RETURN chosen
END

```

```

Communicate Round ( network )
    randomize network's connections
    FOR EACH connection
        diffuse the two agents, once each as the speaker
    END
END

```

```

Diffuse (speaker, hearer)
    pick a random meaning
    pick a random signal from speaker's actives for the meaning
    IF hearer can understand OR speaker can repair
        add speaker id and signal to hearers history
        add signal to passive if not already present
        IF signal already in passive AND chance <= 0.25
            add signal to hearers active
        END
    ELSE
        create new zero level signal
        add signal to each agents active and passive
        add each agents id and signal to each other's history
    END
END

```

```

Can Understand Signal (signal, hearer)
    FOR EACH passive signal hearer has for meaning

```



```

    IF meanings and origins match && renalysis levels are within 1
        RETURN true
    END
END
RETURN false
END

```

```

Can Repair ( speaker, hearer, meaning )
    FOR EACH signal speaker has for the meaning
        IF the hearer can understand
            RETURN true
        END
    END
    RETURN false
END

```

Chapter 3

3.6.1 - 3.6.2

```

Communicate Round ( network )
    randomize network's connections
    FOR EACH connection
        diffuse the two agents, once each as the speaker
    END
END

```

END

END

Diffuse (speaker, hearer)

pick a random meaning

pick a random signal from speaker's actives for the meaning

IF hearer can understand OR speaker can repair

add speaker id and signal to hearers history

add signal to passive if not already present

IF signal already in passive AND chance ≤ 0.25

add signal to hearer's active

END

ELSE

create new zero level signal

add signal to each agents active and passive

add each agent's id and signal to each other's history

END

END

Can Understand Signal (signal, hearer)

FOR EACH passive signal hearer has for meaning

IF meanings and origins match && edit distance between signals ≤ 2

RETURN true

END

END

```
    RETURN false
```

```
END
```

```
Can Repair ( speaker, hearer, meaning )
```

```
    FOR EACH signal speaker has for the meaning
```

```
        IF the hearer can understand
```

```
            RETURN true
```

```
        END
```

```
    END
```

```
    RETURN false
```

```
END
```

```
Segment ( input )
```

```
    segments_store := store for discovered segments
```

```
    partials := partially segmented input, initialized as deep copy of input
```

```
    try and find a segment in partials
```

```
    IF found new segment
```

```
        create segment rule and add to segment_store
```

```
        WHILE can find new segment
```

```
            replace instances of found segment in partials with segment rule symbol
```

```
            try and find a new segment in partials
```

```
        END
```

```
    END
```

```
    finalize segmentation of remaining partials
```

```
RETURN segmented input combined with segment_store
```

```
Find a Segment ( partials )
```

```
data := possible locations for segment
```

```
FOR EACH partial input IN partials
```

```
    FOR EACH contiguous sequence of signal IN the partial input
```

```
        add the signal sequence and the partial input rule's meaning to data
```

```
    END
```

```
END
```

```
starts := count the signs in all signals in data
```

```
frames := initialize a search frames for the three most used signs in starts
```

```
advance search frames until best frame is found
```

```
RETURN found meaning and signal from best search frame
```

```
END
```

```
Initialize Search Frames (signs, data)
```

```
initial_locations := store for initial search frames without meaning
```

```
FOR EACH sign IN signs
```

```
    find each location in data where sign occurs
```

```
    create a search frame from each location and add to initial_locations
```

```
END
```

```
FOR EACH search frame IN initial_locations
```

```

        find all inputs in data where sign occurs and add them to search frame's associated
    END

    initial_morphemes := store for search frames containing all possible meaning signal
    FOR EACH initial_frame IN initial_locations
        FOR EACH meaning IN associated inputs of initial_frame
            create new search frame with sign and meaning
            filter the associated inputs to those that contain both sign and meaning
            add new search frame to initial_morphemes
        END
    END

    RETURN initial_morphemes
END

```

```

Advance Search Frames ( search_frames )
    candidates := store for best search frames, initialized to search_frames
    beam := store for search frames to continue the search, initialized to search_frames
    WHILE still search frames in beam
        advance all search frames
        consider MDL value of advances frames and keep top 5
        update candidates with any potentially better search frames according to MDL and
    END

    RETURN candidates
END

```

Advance Search Frame (search_frame)

IF associated input of search_frame is empty, there are no ways to extend it

RETURN search_frame

END

advances := store for advanced frame

FOR EACH associated input IN search_frame

IF can extend the signal THEN FOR EACH extension

create new search_frame

update signal with extension and keep associated inputs where the updated sign

add new search frame to advances

END

IF can extend the meaning THE FOR EACH extension

create new search_frame

update meaning with extension and keep associated inputs where the updated mea

add new search frame to advances

END

END

RETURN advances

END

Finalize Segmentation of Partial (partials)

FOR EACH partial in partials

IF the partial still has unsegmented meaning and signal

create new segment to with the unsegmented meaning and signals combined

```

        replace occurrence of new segment in partial with new segment signal
        and new segment to segment_store
    END
END
END

```

```

GRIDS ( rewrite_rules )
    WHILE can find new abstract pattern
        update rewrite_rules with pattern
        search for additional patterns in updated rewrite_rules
    END

```

```

    WHILE can find generalization
        update rewrite_rules with generalization
        search for additional generalizations in updated rewrite_rules
    END

```

```

    IF rewrite_rules was updated
        RETURN GRIDS ( _updated rewrite_rules )
    END

```

```

    RETURN ( rewrite_rules )
END

```

Find Pattern (rules)

sequences := store for found patterns

filter out rules that contain phoneme or sememe content

search all rule signals for repeated sequences of length 2, and add to sequences

WHILE found sequences of length n

search for sequences of length n + 1

END

RETURN sequences

END

Find Generalization (rules)

non_starts := filter out rules that begin with start symbol

non_segments := filter out rules that have sememe or phoneme content

FOR EACH rule IN non_starts

FOR EACH other_rule IN non_starts

IF meaning and signal the rules match AND symbols of the rules are different

RETURN symbols of both rules

END

END

END

FOR EACH rule IN non_segments

FOR EACH other rule IN non_segments

IF meanings differ by one sememe AND signals differ by one phoneme in place


```

        RETURN symbols the two symbols that differ
    END
END
END
END

```

```

Update Pattern ( rewrite_rules, patterns)
    sort patterns using MDL and retain top scoring pattern
    create a new rule with pattern symbols as meaning and signal
    replace each occurrence of pattern in rewrite_rules with symbol of new_rule
    RETURN combined updated rewrite_rules with new_rule that captures pattern
END

```

```

Update Generalization ( rewrite_rules, symbols )
    pick one symbol at random to be retained
    replace each occurrence of other symbol with retained symbol in rewrite_rules
    remove one of the now duplicate rules that were used to find the generalization
    RETURN updated rewrite_rules
END

```

```

Transmission ( agent )
    derivations := store for derivations
    active := active repertoire of meaning-signal pairs

```

```

FOR EACH stimuli IN language
    have the agent use their rewrite rules to produce a unique signal for stimuli
    store the derivation of signal in derivations
    add signal to active for that stimuli
END

```

```

FOR EACH segment IN rewrite rules
    count segment symbol in derivation
    IF chance equal to 0.001 times the count
        shorten the segment by one phoneme from end
        reproduce any rules in active that used the segment
    END
END

```

```

IF chance equal to 0.1
    randomly select rule from actives
    randomly remove one sememe from rule if more than one exists
END

```

```

replace with new random signal any signal in actives that have eroded away
set passive repertoire to be a copy of active
remove previous input history
RETURN agent
END

```

Analyze (ARS)

RDG := create rewrite dependency graph from ARS rules

words := store for word-class symbols

syntax := store for syntactic transformation symbols

notsure := store for yet to be analyzed symbols

begin by collecting all symbols in graph pointed to by start symbol

add them to notsure

WHILE not empty notsure

IF symbol does not violates constraints OR symbol is segment, add to words

ELSE add symbol to syntax and add each symbol it can rewrite to to notsure

END

END

RETURN words and syntax

END

Violates Constraints Against Word (symbol)

IF symbol is recursive OR outside rewriteable

RETURN true

END

RETURN false

END

Is Recursive (symbol)

```
    IF endge in graph points to symbol from symbol
        RETURN true
    END
    RETURN false
END
```

```
Is Outside Rewriteable ( symbol, parent )
    IF edge in graph that points to symbol from location other than parent
        RETURN true
    END
    RETURN false
END
```