JOURNAL OF QUANTITATIVE LINGUISTICS
Official journal of the International Quantitative Linguistics Association

Editor
Reinhard Köhler, Department of Computational Linguistics, University of Trier, Trier, Germany

Associate Editors
Gabriel Altmann, Institute of Linguistics, Ruhr Universität, Bochum, Germany
Sheila Embleton, Department of Languages, Literatures & Linguistics, York University, Toronto, Canada
Royal Skousen, Linguistics Department, Brigham Young University, Provo, Utah, USA
Relja Vulanović, Stark Campus, Kent State University, Canton, Ohio, USA

Book Review Editor
Sven Naumann, Department of Computational Linguistics, University of Trier, Trier, Germany

Editorial Board
Michael Cysouw, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
Andrew Hardie, Lancaster University, UK
Luděk Hřebíček, Oriental Institute AS, Prague, Czech Republic
Ján Mačutek, Bratislava, Slovakia
Raimund G. Piotrowski, RPU "Herzen", St. Petersburg, Russia
Pauli Saukkonen, Väylärinne 4, Helsinki, Finland

Journal of Quantitative Linguistics is included in the following abstracting, citation and information retrieval systems: Bibliographie Linguistique/Linguistic Bibliography, Linguistics Abstracts, Linguistics and Language Behavior Abstracts, MLA International Bibliography, Psychological Abstracts, PsycINFO, PsycLIT.

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in its publications. However, Taylor & Francis and its agents and licensors make no representations or warranties whatsoever as to the accuracy, completeness or suitability for any purpose of the Content and disclaim all such representations and warranties whether express or implied to the maximum extent permitted by law. Any views expressed in this publication are the views of the authors and are not the views of Taylor & Francis.

Copyright 2009 by Taylor & Francis

Journal of Quantitative Linguistics
2009, Volume 16, Number 1, pp. 1–39
DOI: 10.1080/09296170802514146

Routledge
Taylor & Francis Group

The Complexity of Discourse*

Paul Kockelman
Columbia University, New York

ABSTRACT

An entropy-based measure is used to calculate the complexity of discourse patterns. This measure of complexity takes into account both the number of patterns (possibly generated), as well as the frequency of each pattern (actually instantiated). The discourse patterns are modelled as random walks in a multi-dimensional and value-weighted space. The multiple dimensions are theoretically specifiable within the framework of conversational analysis (e.g. number of moves, number of participants, types of adjacency pairs, etc.). And the weighted values are empirically measurable within a corpus of texts (e.g. relative frequency participants take the floor, relative frequency first pair-part is a command versus a question, relative frequency an embedding occurs, etc.). The way complexity correlates with various social and discursive factors is described. And the way this method may be extended to analyse successively more complicated patterns is detailed.

1. INTRODUCTION: MEASURING COMPLEXITY AND MODELLING CONVERSATION

An entropy-based measure is used to calculate the complexity of discourse patterns. This measure of complexity takes into account both the number of patterns (possibly generated), as well as the frequency of each pattern (actually instantiated). The discourse patterns are modelled as random walks in a multi-dimensional and value-weighted space. The multiple dimensions are theoretically

*Address correspondence to: Paul Kockelman, Department of Anthropology, Barnard College, Columbia University, 3009 Broadway, New York, NY 10027-6598. Tel: (212) 854-4314. E-mail: pk2113@columbia.edu

0929-6174/09/16010001 © 2009 Taylor & Francis

specifiable within the framework of conversational analysis (e.g. number of moves, number of participants, types of adjacency pairs, etc.). And the weighted values are empirically measurable within a corpus of texts (e.g. relative frequency participants take the floor, relative frequency first pair-part is a command versus a question, relative frequency an embedding occurs, etc.). The way complexity correlates with various social and discursive factors is described. And the way this method may be extended to analyse successively more complicated patterns is detailed.

While the physical nature of entropy is obviously outside of the scope of this essay, the core ideas and fundamental equations go back to the physicist Ludwig Boltzmann (1877; and see Reif, 1965; Tolman, 1979). Linguists familiar with Shannon's classic work (1948) on the mathematical theory of communication should recognize several equations so far as they were originally borrowed from statistical physics. That said, no large-scale metaphysical claims are being made about the entropy (Boltzmann) or information (Shannon) of conversational patterns. Nor will any of the more modern accounts of complexity be taken up, such as the work of Brooks and Wiley (1988), Kauffman (1993), and Simon (1996), *inter alia*. Rather the claim is quite mundane and technical: the mathematical formalism used by physicists to calculate the entropy of thermodynamic systems is a useful way to calculate the complexity of discourse patterns. Usefulness, rather than truthfulness, is the operative term here. No time will be spent arguing that an entropy-based measure is a good measure on logical or *a priori* grounds. Rather, the utility of such a measure will be demonstrated simply by using it. In particular, the calculations undertaken in this essay will show that such a measure is relatively easy to apply, simple to calculate, unambiguous to interpret, and intuitively correct.

While it would be relatively straightforward to write computer programs to model more complicated sequencings of conversational actions (taking into account repair, overlap, openings, and so forth), such an approach has not been taken here. Rather, the central concern has been to obtain, wherever possible, analytic solutions – mathematical equations which are relatively simple to derive and relatively transparent to interpret. The random-walk model of conversation developed here matches these concerns. As will be seen, by slowly adding more and more dimensions to this model, the complexity of more and more elaborate kinds of discourse patterns may be measured – while still maintaining this ethos of derivability and transparency. While the maths itself involves only well-known statistical considerations and algebra, the essay proceeds cumulatively – such that even the least mathematically-inclined linguist should be able to follow all the steps.

In the rest of Section 1, this entropy-based measure of complexity will be introduced. And in Section 2, the random-walk model will be developed, and this measure applied to it.

1.1 Simple Systems

When we measure the complexity of a given system, we are measuring the *degree of freedom* offered by the system: the more formal possibility (i.e. the more states the system could be in), and the more equally distributed these forms are with respect to actual frequency (i.e. the more evenly the system shows up in each of its states), the more complex a system. Loosely speaking, what is a "degree of freedom" from the standpoint of the system is also a "degree of uncertainty" from the standpoint of the observer of that system. That is, complexity is also a measure of the degree of unpredictability of a system: how uncertain we are as to what state it will show up in on any given occasion. To take an example from dicing: assuming a die is unbiased (each side is equally likely to come up), the more sides it has the greater its complexity; and, assuming a die has a particular number of sides, the less biased it is the greater its complexity. These are the essential ideas underlying an *entropy-based measure of complexity*.

To mathematically formulate these core ideas, picture a die with N sides, whose complexity we would like to calculate. Note first that we are not interested in the complexity of the die itself (as a physical object with mass, temperature, volume, etc.). Rather, we are interested in the complexity of its outcomes when rolled. That is, we are interested in the complexity of its probability distribution, as a system which realizes certain states with certain frequencies. Moreover, rather than consider the outcome of a single die, it is best to consider the average outcome of an ensemble of similar dice. In this way, it is best to imagine a very large number of identical dice being rolled at once. Given such a representative ensemble of similar systems, we may ask what percentage of the systems rolled a one, what percentage of the systems rolled a two, and so forth. That is, to calculate the complexity of a given system, we must know the relative frequency (given some ensemble) of each realizable state (given some system).

In particular, if there are N sides of a potentially biased die, whose frequency distribution is given by

$$P = \{P_1, P_2, P_3, \cdots, P_N\} \tag{1.1}$$

where normalization requires that

$$\sum_{i=1}^{N} P_i = P_1 + P_2 + P_3 + \cdots + P_N = 1 \tag{1.2}$$

then the actual complexity of this die, when measured in bits, is given by

$$C_{act} = -\sum_{i=1}^{N} P_i \log_2 P_i. \tag{1.3}$$

The actual complexity of a given system thereby turns on a weighted summation, over possible states, of actual frequencies. *This is the core equation.*

As will be seen, this formulation of complexity has a number of interesting properties. For example, the complexity of a given system is maximum when each formal possibility is equal in frequency. (This should make sense if complexity is a measure of the uncertainty of our knowledge of the state of a system: the more biased a die, the more predictable a die, the less complex a die.) In particular, to say all states are equally frequent is to say that, if there are N states, then each of their frequencies is $1/N$. This means that Equation (1.1) is such that $P_i = 1/N$ for all i. In such a scenario, the maximum complexity of an unbiased die with N sides may be calculated using Equation (1.3), such that

$$C_{max} = -\sum_{i=1}^{N} 1/N \log_2 1/N = -N/N \log 1/N = -\log_2 1/N = \log_2 N. \tag{1.4}$$

That is, the maximum complexity of a system increases logarithmically with the number of states of that system.

As is well known, a logarithm (such as $\log_2 N$) is a very very slowly increasing function of N; it has a zero at $N = 1$; it is negative for $N < 1$; it is undefined for $N < 0$; and it asymptotically approaches negative infinity as N approaches zero. These properties should accord with one's intuitions. For example, it is nearly impossible to imagine a die with less than one side; and, the complexity of a die with only one side would be zero: one knows exactly how the die will turn up on a given occasion. Moreover, given the ever-decreasing slope of the logarithmic function (whose derivative is proportional to $1/N$), the difference in complexity between a die with two sides and a die with three sides is much larger than the difference in complexity between a die with 12 sides and a die with 13 sides. Finally, it should be emphasized that when we say $\log_2 50 = 5.64$, we are saying that $2^{5.64} = 50$, as per the definition of a logarithm, and as per our choice of base 2, such that our unit is the bit. This also emphasizes the fact that a bit is just a unit of measure: we could just as easily measure complexity using decimal digits (base 10), natural units (base e), and so forth. There is nothing inherently binary about complexity anymore than the use of a metre stick means there is something inherently decadic about length.

As may be seen from Equation (1.4), the *maximum complexity* of a given system is only a function of the number of possible states that the system can be in; whereas the *actual complexity*, given by Equation (1.3), is a function of both the number of possible states and the relative frequency with which it is found in each of these states. A useful measure of the *organization* of a system (or the biasing of a die) is the difference between maximum complexity and actual complexity (Brillouin, 1962; Brooks & Wiley, 1988; Layzer, 1988). In particular, we may define the organization of a given system as

$$O = C_{max} - C_{act} = \log_2 N + \sum_{i=1}^{N} P_i \log_2 P_i. \tag{1.5}$$

In other words, to say a system is organized, is to say that its actual complexity does not "tap" its maximum complexity: it is not using its states as equally frequently as it could. Loosely speaking, then, organization is untapped, dispreferred, or prohibited complexity. It usually arises because some external constraints have been imposed on the system (for example, the edges of a die have been filed, or weights have been added to its sides), such that the system does not visit each of its states with equal frequency.

The preceding formulas, and underlying ideas, may be exemplified by appeal to a system of particular interest to linguists: inflectional paradigms consisting of morphologically-encoded meanings. That is, just as a die has a certain number of sides (with a certain degree of relative biasing), a paradigm has a certain number of values (with a certain degree of relative markedness). For example, suppose the system at issue is the simplest person paradigm with three values: first-person, second-person, third-person; and suppose that, in some representative sample of discourse, these values were realized with the following frequencies: 20%, 10%, 70%. Using Equation (1.1), the probability distribution for this system would be

$$P = \{0.2,\ 0.1,\ 0.7\}. \tag{1.6}$$

This is suitable normalized, as per Equation (1.2). Using this distribution in conjunction with Equation (1.3), the actual complexity of this paradigm is given by

$$\begin{aligned} C_{act} &= -\sum_{i=1}^{N} P_i \log_2 P_i \\ &= -0.2\log_2 0.2 - 0.1\log_2 0.1 - 0.7\log_2 0.7 = 1.16\ \textit{bits}. \end{aligned} \tag{1.7}$$

Using Equation (1.4), the maximum complexity of this paradigm is given by

$$C_{max} = \log_2 3 = 1.58\ \textit{bits}. \tag{1.8}$$

Finally, using Equation (1.5), the organization of this paradigm is given by

$$O = C_{max} - C_{act} = 1.58 - 1.16 = 0.42\ \textit{bits}. \tag{1.9}$$

The relevance of this example should be clear. First, just as in the case with dicing, we find the following intuitively correct correlation: assuming that all values in a paradigm are evenly marked (each is equally likely to occur in discourse), the more values a paradigm has the greater its complexity; and, assuming a paradigm has a particular number of values, the more evenly marked these are (as evinced in their relative frequency of usage), the greater its complexity. Second, given the fact that actual complexity turns on both the number of states and the relative frequency of each state, the calculation of linguistic complexity requires that we attend to discursive frequency as much as grammatical structure, *parole* as much as *langue*. And third, as is well known since the work of Zipf (1935) and Greenberg (1966), linguistic forms do not occur with equal frequency because of the relative markedness of the semantic features they encode and the pragmatic functions they serve (as well as the physical properties of the forms which encode). Organization, as the difference between maximum and actual complexity, is therefore a way of measuring the relative markedness of paradigmatic values in regards to semantic features and pragmatic functions – or *value* in the broadest sense. That is, semantic features and pragmatic functions are the linguistic equivalent of constraints on a system. In short, grammatical complexity ultimately turns on the formal encoding of functional specificity (morphosyntax), the frequency of formal tokening (discourse), and the featural and functional mediation of frequency (semantics and pragmatics).

To conclude, it should be stressed that such an entropy-based measure may be applied to any system whose states may be described with a probability distribution. Whether it is a die being rolled in a game, a verb being inflected in an utterance, or a turn being taken in a conversation is immaterial. Moreover, whatever the system at issue, the states are being treated as distinguishable – to the sensory perception of an observer, to the strategies of actors, to the norms of an institution, to the score of a game, etc. This is to highlight the fact that complexity of outcome is ultimately dependent on both the *frame of relevance* and *degree of resolution* with respect to which outcomes are considered. There is no way to account for complexity that is observer-independent, actor-neutral, or meaningfully-autonomous. Ultimately, then, for a comparative approach to complexity, the states must be types – relatively recognizable and repeating tokens, whose relevance and resolution depend on a typology, itself grounded in a particular theory.

In what follows, then, when speaking of a measure of complexity, we mean this entire means – turning on form, frequency, feature, and function; dependent on a frame of relevance and a degree of resolution – of calculating *meaningfully organized complexity*.

1.2 Systems Consisting of Simultaneously Unfolding Subsystems

So far the systems considered have been relatively simple: either one die (with a certain number of sides, which may be more or less biased); or one paradigm (with a certain number of values, which may be more or less marked). With the equations already provided, it is easy to calculate how the complexity of such systems may change as a function of the total number of states and the relative frequency of each state. Before moving on to the complexity of discourse patterns, it is worthwhile calculating the complexity of slightly more complicated systems: first, systems which consist of a number of *simultaneously* unfolding subsystems (such as a roll with more than one die, or constructions with more than one paradigm); and second, systems which consist of a number of *sequentially* unfolding subsystems (such as games with more than one roll, or utterances with more than one construction).

In the case of systems which involve more than one die, the mathematical formalism introduced above may be easily extended. One merely considers a system which consists of several smaller subsystems: for example, all the possible outcomes of M dice, each of which has N sides. To calculate the complexity of such a system, we just count over all the possible states (of the total system), with the assumption that each of these states is equally relevant to the system at issue, the actors implicated in it, or the observers of it. For example, in the case being considered here, we are supposing that the dice are distinguishable (say, one is red and the other green), such that a roll of 1 on one die and 2 on the other is treated as a different outcome from a roll of 2 on the first die and 1 on the second die. In particular, if there are M distinguishable dice, each of which has N sides, then a system consisting of all M dice rolled at once has N^M possible states. This means that the maximum complexity of the total system is given by

$$\begin{aligned} C_{max} &= \log_2 N^M = \log_2(N \times N \times \cdots \times N) \\ &= \log_2 N + \log_2 N + \cdots + \log_2 N = M \log_2 N. \end{aligned} \quad (1.10)$$

That is, the maximum complexity scales linearly with the number of dice in a roll, and logarithmically with the number of sides of each die. And note that Equation (1.10) reduces to Equation (1.5) when $M = 1$. Similarly, in the case of linguistic constructions, maximum complexity increases linearly with the number of paradigms in a construction, and logarithmically with the number of values in a paradigm.

Another way to say all this, as evinced in the middle steps of Equation (1.10), is that the maximum complexity of the sum of the subsystems is equal to the sum of the maximum complexities of the subsystems. This is a key property of complexity, and one which is quite general. In particular, if a system consists of M simultaneously unfolding subsystems, then the maximum complexity of this system is given by

$$C_{max}^{1+2+3+\cdots+M} = C_{max}^{1} + C_{max}^{2} + C_{max}^{3} + \cdots + C_{max}^{M}. \quad (1.11)$$

And, as was implicit in the definition of organization given in Equation (1.5), the actual complexity of a system is always less than or equal to its maximum complexity. That is, the following inequality holds, essentially by definition:

$$C_{act} \leq C_{max}. \quad (1.12)$$

Now just as the sides of a single die may be *biased* (or the values of a single paradigm may be *marked*), the sides of simultaneously rolled dice may be *correlated* (or the values of co-occurring paradigms may be *dependent*). For example, not only does first-person have a different frequency than second-person and third-person (hence, the system exhibits markedness), but the relative frequency of first-person is different depending on the current value of co-occurring paradigms – such as whether number is singular or plural, or whether tense is present or past (hence the system exhibits dependency). Loosely speaking, and in a Saussurian idiom, markedness is a constraint imposed upon an axis of selection (which increases paradigmatic organization); dependency is a constraint imposed upon an axis of combination (which increases syntagmatic organization). Just as the relative markedness of values within a single paradigm reduces the actual complexity of the paradigm, as per Equation (1.12), the relative dependence between values of co-occurring paradigms reduces the actual complexity of the construction. In other words, organization increases not just with increased biasing of *intra*-subsystem values, but also with increased correlation of *inter*-subsystem values.

This means that a more general set of relations holds. Firstly, the generalization of Equation (1.11) may be expressed as follows:

$$C_{act}^{1+2+\cdots+M\ (uncorrelated)} = C_{act}^{1} + C_{act}^{2} + \cdots + C_{act}^{M}. \quad (1.13)$$

That is, when the simultaneously unfolding subsystems of a system are uncorrelated (with each other) but potentially biased (within themselves), the actual complexity of the system is equal to the sum of the complexities of the subsystems; and secondly, when the simultaneously unfolding subsystems of a system are potentially correlated (with each other) and potentially biased (within themselves), the actual complexity is usually less than the actual complexity of the same subsystems that are uncorrelated. In other words,

$$C_{act}^{1+2+\cdots+M\ (correlated)} \leq C_{act}^{1+2+\cdots+M\ (uncorrelated)}. \tag{1.14}$$

In short, biasing and correlation (or markedness and dependency, in the case of grammatical constructions) reduce the actual complexity of a system from its maximum complexity. This should come as no surprise: biasing and correlating are ways of organizing a system, and they count as constraints that have been imposed upon a system. The more constraints on a system, the more organized a system, the more predictable a system, the less complex a system.

1.3 Systems Consisting of Sequentially Unfolding Subsystems

Having analysed the complexity of a single die (with a certain number of sides and a certain degree of biasing), and having analysed the complexity of a system of dice (with a certain number of dice and a certain degree of correlation), we may now turn to the complexity of a certain number of rolls of a die (or rolls of a system of dice). That is, we may turn to systems consisting of sequentially unfolding subsystems. To take the simplest example – one which has incredibly broad implications – consider a die with two sides, or what is in essence a coin. To review: if this coin were unbiased, then the maximum complexity of a single flip would be $\log_2 2$, or 1 bit, following Equation (1.4). Were this coin biased, such that heads had a relative frequency of p, and tails had a relative frequency of $1-p$ (where $p \neq 1/2$), then the complexity of a single flip would be given by Equation (1.3), and would lie somewhere between 0 bits and 1 bit, depending on the amount of biasing. Were we to flip several unbiased coins, then the complexity of the distribution would be M bits, where M is the number of coins flipped, following Equation (1.10), and so forth. All this is easily understood given the preceding discussion. What is at issue now, however, is the complexity of a system

which consists of many flips of a single coin (or, more generally, many rolls of a set of dice).

Consider N tosses of a biased coin. With each toss, the coin has a probability p of coming up heads and a probability $1-p$ of coming up tails (where $p=1-p=1/2$ in the case of an unbiased coin). After a total of N tosses, we would like to know the probability that n_1 of the tosses came up heads, and $n_2 = N - n_1$ of the tosses came up tails. Analogously, consider N steps of a tipsy sailor. With each step the sailor has a probability p of going to the right, and a probability $1-p$ of going to the left. After a total of N steps, we would like to know the probability that n_1 of the steps were to the right, and $n_2 = N - n_1$ of the steps were to the left. The answer is the famous *binomial distribution* (see Reif, 1965, for a derivation):

$$W_N(n_1) = \binom{N}{n_1} p^{n_1}(1-p)^{N-n_1} \tag{1.15}$$

where

$$\binom{N}{n_1} = \frac{N!}{n_1!(N-n_1)!}. \tag{1.16}$$

As a probability distribution, it is suitable normalized, such that

$$\sum_{n_1=0}^{N} W_N(n_1) = \sum_{n_1=0}^{N} \binom{N}{n_1} p^{n_1}(1-p)^{N-n_1} = 1. \tag{1.17}$$

Moreover, it has an *average*, given by

$$\overline{n_1} = \sum_{n_1=0}^{N} W_N(n_1)\, n_1 = \sum_{n_1=0}^{N} \binom{N}{n_1} p^{n_1}(1-p)^{N-n_1}\, n_1 = Np. \tag{1.18}$$

And it has a *dispersion*, given by

$$\overline{(\Delta n_1)^2} = \overline{(n_1 - \overline{n_1})^2} = Np(1-p). \tag{1.19}$$

As will be taken up below, this distribution may often be approximated by a Gaussian (or bell-shaped curve): the average corresponds to the location of the maximum of the distribution; and the square-root of the dispersion (or root-mean-square deviation) is a good measure of the width of the distribution around the maximum point.

As with any probability distribution, we may calculate the complexity of a binomial distribution with the usual formula. In particular, following Equation (1.3), we find

$$C_N(p) = -\sum_{n_1=0}^{N} W_N(n_1)\log_2 W_N(n_1)$$
$$= -\sum_{n_1=0}^{N}\binom{N}{n_1}p^{n_1}(1-p)^{N-n_1}\log_2\binom{N}{n_1}p^{n_1}(1-p)^{N-n_1}. \quad (1.20)$$

Equation (1.20) is graphed for a range of values in Figure 1.1. As may be seen, complexity is a slowly increasing function of N. For a given N, complexity is maximum when $p = 1 - p = 1/2$. And complexity is symmetric with respect to p and $1 - p$: that is, the complexity of

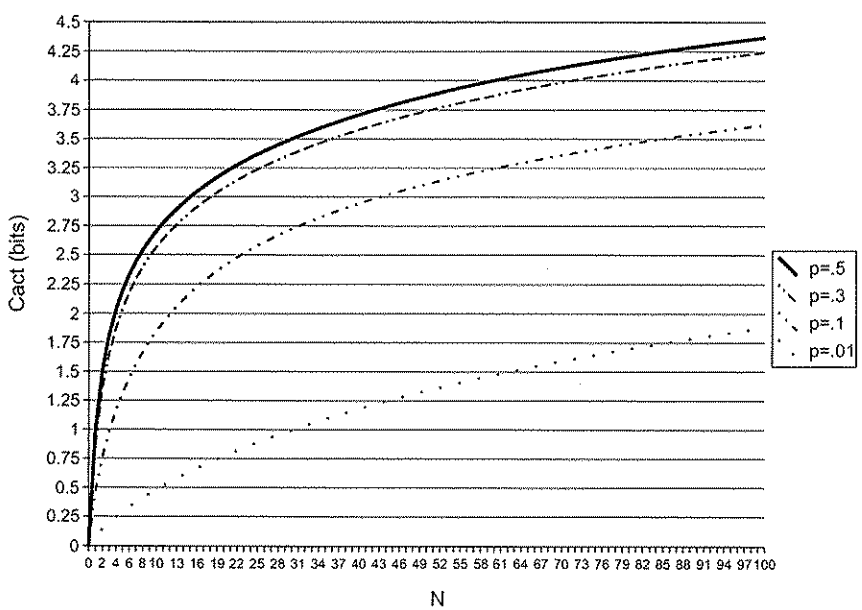

Fig. 1.1. Complexity of binomial distribution.

a distribution with $p = 1/3$ and $1 - p = 2/3$ is the same as the complexity of a distribution with $p = 2/3$ and $1 - p = 1/3$. In other words, as long as the overall biasing is the same, whether it is biased towards tails or heads is immaterial.

If one were to move from a system consisting of successive flips of a biased coin to a system consisting of successive steps of a tipsy sailor, Equation (1.20) is a measure of the relative complexity of the *random walk* taken by the tipsy sailor. For example, the more steps a tipsy sailor takes, the more complex the overall path taken. And a tipsy sailor dead-set on moving to the right (such that $p \approx 1$) is much less complex than one who is equally likely to go to the right or the left (such that $p \approx 1 - p \approx 1/2$). Indeed, if we think of increasing sobriety as leading a sailor to move in one direction (p approaches 1), and tipsiness as leading a sailor to move in either direction (p approaches 1/2), then a tipsy sailor is a complex sailor (whose behaviour is relatively unpredictable), whereas a sober sailor is a simple sailor (whose behaviour is relatively predictable). Or, from the standpoint of organization, which is equal to the difference between maximum complexity and actual complexity, the path of a tipsy sailor is relatively disorganized, and the path of a sober sailor is relatively organized.

As is well known (see Reif, 1965, for a detailed derivation), in the limit of large N (when p and $1 - p$ are not too small), the binomial distribution may be approximated by a *Gaussian distribution*. In particular, one finds that Equation (1.15) may be approximated by

$$W_N(n_1) \approx W(N, n_1) = \frac{\exp^{\frac{-(n_1-Np)^2}{2Np(1-p)}}}{\sqrt{2\pi Np(1-p)}},\ \textit{for}\ Np(1-p) \gg 1. \quad (1.21)$$

Just as we may approximate a binomial distribution by a Gaussian distribution (and hence by a continuous function), we may also calculate the complexity of a Gaussian distribution, which is a good approximation of the complexity of a binomial distribution in the limit of large N (when p and $1 - p$ are not too small). As derived in Shannon (1963, pp. 88–89), for the case of "information", this involves an integral rather than a sum, and yields

$$C_N(p) \approx C(N,p) = \log_2\sqrt{2\pi eNp(1-p)},\ \textit{for}\ Np(1-p) \gg 1. \quad (1.22)$$

This, then, is an analytic solution to the complexity of a random walk, and hence a useful approximation of Equation (1.20). As per the discussion following Equation (1.20), but now with analytic detail, the complexity of a random walk increases with the logarithm of the square root of N; is maximum, for a given N, when $p = 1 - p = 1/2$; and is symmetric with respect to p and $1 - p$. Indeed, as may be seen by comparing Equation (1.19) with Equation (1.22), *the complexity of a Gaussian approximation to a binomial distribution is essentially the logarithm of the root-mean-square deviation*. Thus, the greater the width of a distribution, the greater the complexity of the distribution: an extremely narrow distribution only has a significant probability for a small number of states, paths, or outcomes – and hence is maximally organized and minimally complex. The path of a sober sailor, and thus an organized sailor, is essentially known in advance.

These then are the essential features of the complexity of sequentially unfolding subsystems, as exemplified by successive flips of a biased coin or successive steps of a tipsy sailor. Two additions should be made before concluding this section. First, so far we have been focused on sequentially unfolding subsystems which themselves have only two states (heads or tails, right or left). The reasoning leading to Equation (1.15) is readily generalized for sequentially unfolding subsystems of any number of states. For example, rather than flip a coin we might roll a three-sided die. In such a scenario, the probability distribution is given by

$$W_N(n_1, n_2) = \frac{N!}{n_1!n_2!(N - n_1 - n_2)!} p_1^{n_1} p_2^{n_2} (1 - p_1 - p_2)^{N - n_1 - n_2}. \quad (1.23)$$

This is the probability that, out of a total of N rolls, n_1 rolls come up 1, n_2 rolls come up 2, and $N - n_1 - n_2$ rolls come up 3, where the probability of rolling a 1 is given by p_1, the probability of rolling a 2 is given by p_2, and the probability of rolling a 3 is given by $1 - p_1 - p_2$. This is suitably normalized such that

$$\sum_{n_1=0}^{N} \sum_{n_2=0}^{N-n_1} W_N(n_1, n_2) = 1. \quad (1.24)$$

Moreover, it too may be approximated by a Gaussian; and it too may have its complexity calculated. Such results, and their generalization (for, say, N rolls of a four-sided die, N rolls of a five-sided die, and so forth), will come in handy in the next section.

Second, we may focus on the complexity of sequentially unfolding subsystems which themselves consist of simultaneously unfolding subsystems. That is, we may bring the results of this section to bear on the results of the preceding section. Imagine, for example, that a number of different-sided dice are thrown with every turn. Or, imagine that a grammatical construction involving five different paradigms (say, all operators with scope over verbal predicates: illocutionary force, tense, mood, polarity, aspect) is successively instantiated a certain number of times in a given narrative. Indeed, to take the most relevant example for modelling conversation, imagine three coins are flipped with every turn. Or, equivalently, but couched in terms of human action, imagine that the random walk of a sailor takes place in three dimensions (north versus south, east versus west, and up versus down). Or, even more picturesquely, imagine that, with each step a tipsy sailor has three choices to make, and hence three kinds of actions to simultaneously undertake: whether to move to the right or left, whether to adjust his cap or tighten his belt, and whether to sing or shout. Given the fact that independent probabilities multiply (and the fact that, for the moment, we are treating the simultaneously undertaken actions as uncorrelated), the probability distribution of such a system is given by

$$\begin{aligned} &W_N^{1+2+3}(n_1, n_2, n_3) \\ &= W_N^1(n_1) W_N^2(n_2) W_N^3(n_3) \\ &= \binom{N}{n_1} p^{n_1} (1 - p)^{N - n_1} \binom{N}{n_2} q^{n_2} (1 - q)^{N - n_2} \binom{N}{n_3} r^{n_3} (1 - r)^{N - n_3}. \quad (1.25) \end{aligned}$$

That is, Equation (1.25) shows the probability that, out of N flips of three coins, the first coin comes up heads n_1 times (the sailor moves to the right), the second coin comes up heads n_2 times (the sailor adjusts his cap), and the third coin comes up heads n_3 times (the sailor sings), where p is the probability that the first coin comes up heads, q is the probability that the second coin comes up heads, and r is the probability that the third coin comes up heads. It is suitably normalized, such that

$$\sum_{n_1=0}^{N} \sum_{n_2=0}^{N} \sum_{n_3=0}^{N} W_N^{1+2+3}(n_1, n_2, n_3) = 1. \quad (1.26)$$

Equation (1.25) is easily approximated using a Gaussian distribution. Indeed, it is just the product of three equations similar to Equation (1.21), and the complexity of such a Gaussian distribution is easily calculated. Indeed, as per the discussion regarding Equation (1.13), just as the complexity of a grammatical construction involving a set of independent paradigms is just the sum of the complexities of the independent paradigms, the complexity of a random walk in three (uncoupled) dimensions is just the sum of the complexities of each individual dimension. In particular, in the limit that $Np(1-p) \gg 1$, $Nq(1-q) \gg 1$, and $Nr(1-r) \gg 1$, we find

$$C(N,p,q,r) = \log_2 \sqrt{2\pi e N p(1-p)} + \log_2 \sqrt{2\pi e N q(1-q)} + \log_2 \sqrt{2\pi e N r(1-r)}. \quad (1.27)$$

Equations (1.25)–(1.27) may be generalized for random walks in any number of dimensions. Moreover, they may be used in conjunction with the generalization of Equation (1.23), such that these simultaneously unfolding subsystems may themselves each have any number of states.

Having detailed a method of calculating complexity that is applicable to a wide variety of systems, we may now turn from the complexity of solitary sailors taking steps home from a bar to the complexity of social speakers taking turns in a conversation. As will be seen, the discursive patterns evinced in such conversations are easily modelled as random walks in coupled, multi-dimensional, value-weighted spaces.

2. THE COMPLEXITY OF DISCOURSE: APPLYING THE MEASURE TO A MODEL

In this section a model of discourse is used which loosely follows the ideas of conversational analysis (cf. Sachs et al., 1974; Goffman, 1981; Schegloff, 2005). The basic tact is to use a model of discourse patterns that is highly constrained, such that the calculations are relatively tractable, and the logic relatively transparent; yet, at the same time to use a model of discourse patterns that is easily generalized, such that the formalism may be readily extended to more elaborate and empirically realistic patterns. As will be seen, the best way to do this is to model discourse as a *random walk*. Within such a broad metaphor, the number of moves made in a conversation (or any sequence of moves within a conversation) is analogous to the number of steps taken in a walk. The number of participants and adjacency pairs is analogous to the dimensions in which one may step. And the relative frequency that one participant takes the floor (rather than another), or that one adjacency pair is used (rather than another), is analogous to the probability that a step is to the right or to the left. Depending on the kind of conversation to be modelled, such properties may be varied, added, or coupled, in increasingly complicated, yet analytically tractable ways.

In the most basic example, we will calculate the complexity of a conversational sequence of N moves, undertaken by two participants, when there are two types of adjacency pairs (say, question-answer and command-undertaking). One might imagine the simplest kind of language game undertaken by a diner and a waiter. To these basic parameters (number of moves, number of participants, number of adjacency pairs), there are three key variables.

First, whenever the floor is open, there is the probability p that participant 1 makes a move versus the probability $1-p$ that participant 2 makes a move.

Second, whenever the move at issue is a first pair-part, there is the probability q that such a move is a question versus the probability $1-q$ that such a move is a command.

Third, whenever the move at issue is a response to a first pair-part, there is the probability ε that it is the preferred second pair-part (for example, a question is followed by an answer) versus the probability $1-\varepsilon$ that an embedding occurs, such that another adjacency pair is inserted (for example, a question is followed by a question).

As an example of a sequence with four moves, involving two participants, two adjacency pairs (both of which are question-answer couplets), and one embedding, we may turn to the classic paper by Sachs et al. (1974, p. 702):

Anna: Was last night the first time you met Missiz Kelly?
Bea: Met whom?
Anna: Missiz Kelly.
Bea: Yes.

Using the entropy-based measure outlined in Section 1, not only can we see how the complexity of discourse varies as a function of the *types of* *parameters* (number of moves, participants, and adjacency pairs) but, given a set of parameters, we can see how the complexity of discourse varies as a function of the *values of the variables* (p, q, ε). Moreover, as linguists, psychologists, and anthropologists, we may see how the parameters and variables depend on various properties of communicative systems, cognitive processes, and social contexts. This opens a large-scale empirical and comparative project, whose overarching goal is to examine the conditions for, and consequences of, various modes of meaningfully organized complexity.

Before undertaking the analysis, a number of readily intelligible disclaimers should be made. First, when speaking of the complexity of a conversation, no attention is being paid to the content, size, or medium of a given move. All that counts are the types of moves made, from the standpoint of the kinds of actions undertaken: question and answer, command and undertaking, offer and acceptance, announcement and assessment, complaint and remedy, etc. The actual content (what is said or done), length (how long it takes to say or do), or medium (be it an utterance, gesture, facial expression, or physical action) is not at issue. The approach outlined here is thereby designed to apply to any form of interaction – sports, games, dances, markets. Nevertheless, the account of grammatical constructions outlined in the previous sections (especially Section 1.2) may easily be extended to include lexical constructions; and such an account of lexico-grammatical constructions may be combined with this account of discourse patterns. For example, to join this analysis with that of scholars interested in the relation between discourse and grammar (Du Bois, 2003; Sachs et al., 1974, p. 702; Ochs et al., 1996; *inter alia*), one could examine the ways various turn-constructional units (be they sentential, clausal, phrasal, lexical) are used to constitute moves, and the way various grammatical categories (person, number, case, etc.) and lexical categories (nouns and verbs) are used to constitute turn-constructional units. In this way, a more extended analysis can be used to account for the meaningful content of the moves at issue.

Second, just as moves may be positioned relative to other moves, as part of a coherent sequence of action, moves may also be positioned relative to a stretch of talk itself, such as opening and closing brackets, greetings and goodbyes, and so forth (Duranti, 1997; Schegloff & Sachs, 1973; Schegloff, 1986). While these are not treated here, they are often the most organized part of a conversation, and the analysis could easily be extended to include them. Indeed, as will be taken up in Section 2.5, while the analysis here is focused on the ordering of moves into coherent sequences, one could use similar techniques to account for the ordering of sequences into coherent conversations. This is a crucial point: just as the analysis may be extended "downwards" into the grammatical and lexical content of moves, the analysis may also be extended "upwards" into the patterning of sequence types into conversations.

Third, in this model a turn – as a single stretch of talk in which a particular participant makes one or more moves – is not an independent variable (cf. Schegloff, 1996). However, how many turns there are in a conversation, and how many moves such turns consist of, may be calculated as a statistical property (in terms of averages and dispersions) from the other variables considered here. For example, if one knows the probability that a participant makes a specific number of moves in a conversation of a particular length, one has information about the average length (in terms of moves) of the participant's turns – with the assumption that sequential moves by the same speaker constitute a single turn. More generally, there is all sorts of statistical information that could be calculated (given the probability distributions to be derived), but which will not be considered (given the focus of this essay).

Fourth, just as simultaneously unfolding subsystems may be correlated, so may sequentially unfolding subsystems. For example, the probability that a tipsy sailor moves to the right or the left may be a function of the last move the sailor made (or the last two moves, or the last three moves, etc.). For present purposes, such correlations are incorporated insofar as first pair-parts condition second pair-parts, no matter how tenuously conditioned or delayed. More complicated kinds of correlations, requiring mathematical analysis of Markovian processes, will not be treated here.

Fifth, the other range of issues for which conversational analysis has been so successful – such as the repair of mistakes and the overlap of turns – are not treated here at all (cf. Jefferson, 1984; Schegloff, Jefferson & Sachs, 1977); though certain aspects of repair, as evinced in the discourse example given above, are readily tractable via embedded sequences. Moreover, while there are a range of moves which do not make responses relevant, and hence fall out of an adjacency-pair format (Schegloff, 2004, pp. 10–11), these are not treated here either. The techniques described in Section 2.2, however, may readily be used to incorporate them.

That said, conversational analysis has had a distinguished career for over 30 years, providing a theoretically sophisticated and empirically rich approach to the analysis of conversation. From its beginnings, it has been interested in the relation between context-sensitive and context-free forms of organization, turning on distinctions similar to parameters and variables. And, in the tradition of George Herbert Mead and Erving Goffman, it has been particularly successful in modelling such patterns of sequentially coherent actions. As Schegloff (2004, p. 8), one of its forefathers, stresses:

> If we ask how actions and courses of actions get organized in talk in interaction, it turns out that there are a few kernel forms of organization that appear to supply the formal framework within which the context-specific actual actions and trajectories of actions are shaped. By far the most common and consequential is the one we call "adjacency pair based".... The simplest and minimal form of a sequence is two turns long: the first *initiating* some kind of action trajectory – such as requesting, complaining, announcing, and the like, the second responding to that action in either a compliant or aligning way (granting, remedying, assessing and the like, respectively) or in a disaligning or non-compliant way (rejecting, disagreeing, claiming prior knowledge and the like, respectively).

It is precisely these adjacency pairs – as *the most common and consequential kernel forms of organization* – that are at issue in the following sections.

2.1 The Simplest Model: Two Participants, Two Adjacency Pairs

For the moment, set aside the issue of who makes a move (participant 1 versus participant 2); and set aside the issue of which kind of move is made (question versus command). Focus instead on whether a move is a first pair-part (either a question or a command) versus a second pair-part (either an answer or an undertaking). Assuming that no sequence may end until all first pair-parts have been responded to by second pair-parts (that is, all questions are answered and all commands are undertaken), a given sequence can only have an even number of moves: $N=2$, 4, 6, 8, etc. This is the first key constraint – and it is built directly into the model. The second key constraint is even more basic: as their name implies, first pair-parts always precede second pair-parts (either immediately, or with some delay). That is, an answer to a question never precedes the question; and an undertaking of a command never precedes the command.

Notice that neither of these constraints was present in the random-walk model introduced in Section 1.3. They are equivalent to a random walk that must always take an even number of steps, and a random walk in which no steps to the left can occur unless there has already been at least as many steps to the right. More graphically, they are equivalent to a random walk that must always return to the origin, and a random walk that never strays to the left of the origin. The only issue that is left to chance (or "choice" as the case may be), and hence the only open variable, is whether a second-pair part immediately follows a first pair-part (e.g. a question is immediately responded to with an answer), or whether *embedding* occurs, such that a first pair-part is responded to by another first pair-part (e.g. a question is responded to with a question, and is only answered after this intermediate question has been answered). When such an embedding occurs – and there may be more than one such embedding in a row – there is a delay, consisting of some number of moves, between a first pair-part and its second pair-part. In the case of an actual random walk (which must always return to the origin, and which can never stray to the left of the origin), this is equivalent to asking whether a step to the left immediately follows a step to the right, or whether some number of steps to the right can intervene.

For example, in a sequence of two moves (when, for the moment, we are still not considering the number of participants or the types of adjacency pairs), there is only one "path" through the space of possible moves: the first move is a first pair-part and the second move is a second pair-part (in response to the first move). For a conversation of four moves, there are two paths: either there is no embedding such that the first move is first pair-part, the second move is a second pair-part (in response to the first move), the third move is a first pair-part, and the fourth move is a second pair-part (in response to the third move); or there is a single embedding such that the first move is a first pair-part, the second move is a first pair-part, the third move is a second pair-part (in response to the third move), and the fourth move is a second pair-part (in response to the first move). Figure 2.1 diagrams these results for $N=2$, 4, 6, 8. As may be seen, such paths always return to the origin (all first pair-parts are responded to by second pair-parts), and never cross the origin (second pair-parts are always responses to first pair-parts). (The boxed portion of Figure 2.1 [where $N=6$ and $N=8$] shows a type of

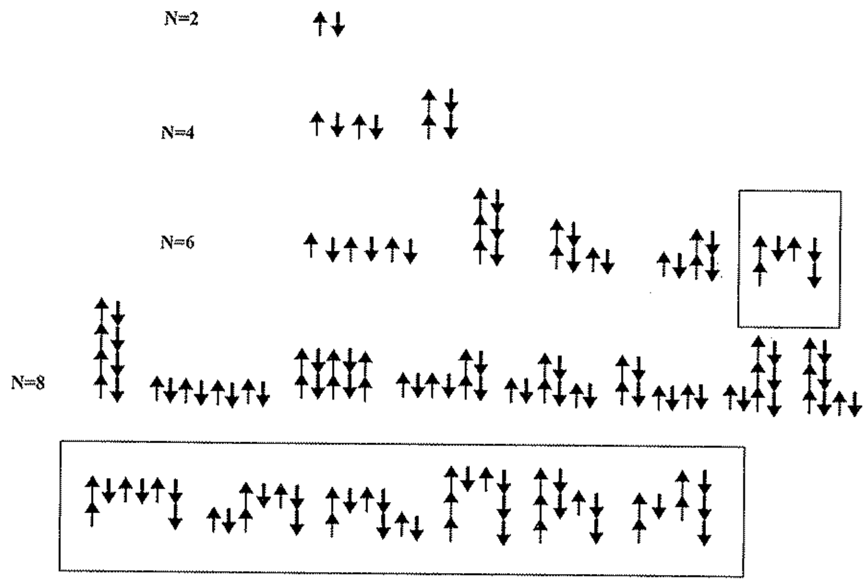

Fig. 2.1. Types of embeddings for sequences of different lengths.

path that will not be considered here: the unfolding of moves at a particular level of embedding.)

As may be intuited from these diagrams, the embedding dimension of a sequence has a random-walk structure, which is subject to two constraints. Its distribution is given by the following equation

$$W_N(m) = \binom{N/2 - 1}{m} \varepsilon^m (1 - \varepsilon)^{N/2-1-m} \quad (2.1)$$

which is normalized, such that

$$\sum_{m=0}^{N/2-1} W_N(m) = \sum_{m=0}^{N/2-1} W_N(m) = \binom{N/2 - 1}{m} \varepsilon^m (1 - \varepsilon)^{N/2-1-m} = 1. \quad (2.2)$$

Compare Equation (1.15) and Equation (1.17). Here N denotes the number of moves in the sequence (and may range from 2 to infinity, counting even numbers only), ε denotes the probability that an embedding occurs (and ranges from 0 to 1), and m denotes the number

of embeddings that may occur in a sequence of N moves (and ranges from 0 to $N/2 - 1$).

For example, when $N = 2$, $m = 0$ and $W_N(m) = 1$ (there is no possibility of embedding for a sequence with only two moves). When $N = 4$, $m = 0$ and $W_N(m) = 1 - \varepsilon$, or $m = 1$ and $W_N(m) = \varepsilon$, (one type of sequence with four moves has no embedding, and another type has one embedding). When $N = 6$, $m = 0$ and $W_N(m) = (1 - \varepsilon)^2$, $m = 1$ and $W_N(m) = 2\,(1 - \varepsilon)$, or $m = 2$ and $W_N(m) = \varepsilon^2$ (one type of sequence with six moves has no embedding, another type has one embedding, and another type has two embeddings); and so forth. Recall Figure 2.1. Note that each of these is normalized, and so when one adds up all the probabilities of different types of sequence (of a given length), the result is unity. And note that, when $\varepsilon \ll 1$, any probabilities which involve higher and higher order powers of ε (such as ε^2, ε^3, etc.) become vanishingly small. Thus, the fact that embedding may occur to any degree (given the types of parameters) does not mean that multiple embeddings are necessarily likely (given the values of the variables). The presence of the factor $N/2$ instead of N is due to the constraint that we are only considering sequences with an even number of moves; and the presence of the factor $N/2 - 1$ is due to the fact that the first move of any sequence is necessarily a first pair-part (neither responding, nor embedding, is an option). Otherwise Equation (2.1) is identical to Equation (1.15).

Finally, as already intimated, one suspects that the variable ε is relatively small for most sequences: that is, sequences are relatively organized with respect to this parameter, as evinced by their being "flat" rather than "embedded". However, as a function of the sequence in question, and the type of conversation in which the sequences occurs, the degree of preference may be altered by adjusting the value of ε. The smaller the value of the variable ε, the more a sequence is organized with respect to this parameter, the less likely a sequence is to embed, and hence the less likely a sequence will exhibit the "higher-order" patterns depicted in Figure 2.1.

Before calculating the complexity of a sequence described by this probability distribution, we need to include the two parameters (and variables) left out of the preceding discussion: the probability that one participant takes the floor, rather than another (where there are two participants); and the probability that a given move is a question rather than a command (where there are two types of adjacency pairs). Assuming that neither of these parameters is dependent on the other, nor on the degree of embedding, they can each be modelled as random walks. In particular, anytime the floor is open, one participant or the other can take it, and anytime a move is made (which is a first pair-part), it can be a question or a command. Here then we are in the realm of the tipsy sailor who has to make three choices with every step: whether to go to the left or the right (embed or not), whether to adjust his cap or tighten his belt (participant 1 or participant 2), and whether to sing or shout (question or command). Incorporating Equation (2.1), and with analogy to Equation (1.25), the probability distribution that describes such sequences is

$$W_N(m,r,s) = \binom{N/2-1}{m}\varepsilon^m(1-\varepsilon)^{N/2-1-m}\binom{N/2}{r} \times p^r(1-p)^{N/2-r}\binom{N/2}{s}q^s(1-q)^{N/2-s} \qquad (2.3)$$

which is suitably normalized, such that

$$\sum_{m=0}^{N/2-1}\sum_{r=0}^{N/2}\sum_{s=0}^{N/2} W_N(m,r,s) = 1. \qquad (2.4)$$

Here ε denotes the probability that a move is embedded (as before), p denotes the probability that the move is made by participant 1 (versus participant 2), and q denotes the probability that the move is a question (versus a command). The summations (over participants and adjacency pairs) also only goes up to $N/2$ because, by the constraints built into the model, only first pair-parts and floor-takers are open to chance (or choice); second pair-parts and addressees (regardless of how long they are deferred due to embedding), are always ultimately determined by first pair-parts and speakers. It should be emphasized that Equation (2.3) denotes the probability that a sequence of N moves will have m embedded moves (and $N/2-1-m$ non-embedded moves), r moves initiated by participant 1 (and $N/2-r$ moves initiated by participant 2), and s questions (and $N/2-s$ commands). It does not tell us where the embeddings occur, or which moves are made by participant 1, or which moves are questions. Averages and dispersions, as well as any other statistical information, may be calculated from this distribution using the analogues of Equation (1.18) and Equation (1.19).

With one caveat (regarding the unfolding of moves at a particular level of embedding, as per the boxed portions of Figure 2.1), the discourse patterns modelled by Equation (2.3) may be geometrically represented by Figure 2.2. In particular, any sequence starts off at the centre point. The first move made may be a question or a command, and it may be made by participant 1 or participant 2. Hence, there are four paths that may be taken from the centre point (to a nearby radial point), each with a different probability depending on the values of p and q. If the sequence consists of only two moves ($N=2$), then the second move must be a return to the centre point from one of the first proximal radial points (of which there are four): a question asked by participant 1 is answered by participant 2, a command given by participant 2 is undertaken by participant 1, and so forth. If the sequence consists of only four moves, the first move is just as before: say, a question is asked by participant 1. The second move may either be an answer to that question by participant 2, or it may be a question or command by participant 2 (thus an embedding with a change of speaker), or it may be a question or command by participant 1 (thus, an embedding without a change in speaker). In this way, the second move may either return the conversation to the centre point, or it may radiate out again to a second-most proximal radial point (of which there are 16, in total, rather than four). If it returns to the centre, then it can repeat itself: one participant or the other takes the floor, and the move made is either a question or command; and, if it moves out to a

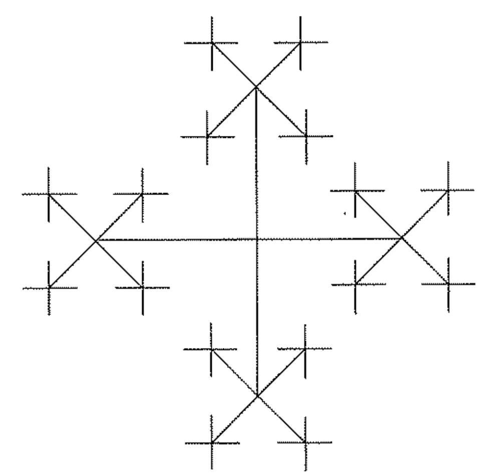

Fig. 2.2. Fractal embedding of conversational patterning.

more radial point, given the fact that the conversation is only four moves long, the next two moves must bring it back to the centre point again: the second first-pair part is given a response; and then the first pair-part is given a response.

With the addition of more types of adjacency pairs, Figure 2.2 may be extended by adding more lines radiating out from any point. That is, by changing the number of parameters, one changes the number of "spokes". For example, if there were three types of adjacency pairs rather than two, each point would have six spokes radiating out from it. Moreover, with the tweaking of probabilities, the types of paths taken through this space tend to cluster in one portion or another. That is, by changing the value of the variables, one changes the amount of "skew". For example if p, q and ε were all nearly 0, then most sequences would take place on one spoke, in between the centre point and the first most proximal radial point. Qualitatively speaking, the more spokes, the greater the maximum complexity of a sequence; and the more skewing, the greater the organization of a sequence. A sequence whose actual complexity is close to its maximum complexity, and hence one which is minimally organized, is one in which the entire space allotted is readily filled.

As may be seen by Equation (2.3), this distribution is given by the product of three binomial distributions (which should make sense: the three parameters are independent of each other, or uncoupled). The complexity of such a distribution, by analogy with Equation (1.20), is given by

$$C_N(\varepsilon, p, q) = \sum_{m=0}^{N/2-1} \sum_{r=0}^{N/2} \sum_{s=0}^{N/2} W_N(m, r, s) \log_2 W_N(m, r, s). \tag{2.5}$$

While one may calculate these values directly (for $\varepsilon = p = q = 1/2$, $1/10$, $1/100$), it is worthwhile using the Gaussian approximation in order to obtain an analytic solution. Thus, in analogy with Equation (1.21), in the limit of $(N/2-1)\ \varepsilon\ (1-\varepsilon) \gg 1$, $(N/2)\ p\ (1-p) \gg 1$, and $(N/2)\ q\ (1-q) \gg 1$, Equation (2.3) may be approximated by

$$W(N, m, r, s) = \frac{\exp \frac{-(m-(N/2-1)\varepsilon)^2}{2(N/2-1)\varepsilon(1-\varepsilon)}}{\sqrt{2\pi(N/2-1)\varepsilon(1-\varepsilon)}} \frac{\exp \frac{-(r-(N/2)p)^2}{2(N/2)p(1-p)}}{\sqrt{2\pi(N/2)p(1-p)}} \times \frac{\exp \frac{-(s-(N/2)q)^2}{2(N/2)q(1-q)}}{\sqrt{2\pi(N/2)q(1-q)}}. \tag{2.6}$$

Finally, we may use this Gaussian approximation to obtain an analytic result which describes the complexity of discourse patterns generated by this model, and hence approximates the solution to Equation (2.5). In analogy with Equation (1.27), we arrive at

$$C(N, \varepsilon, p, q) = \log_2 \sqrt{2\pi e(N/2-1)\varepsilon(1-\varepsilon)} \sqrt{2\pi e(N/2)p(1-p)} \times \sqrt{2\pi e(N/2)q(1-q)}. \tag{2.7}$$

This is an elegant result: the complexity of this distribution is essentially equal to the logarithm of the product of the root-mean-square deviations of the three Gaussians whose product makes up the distribution.

More generally, given the fact that $\log(A \times B \times C) = \log(A) + \log(B) + \log(C)$, this shows that the complexity of the sequence is simply the sum of the complexities of each separate dimension (embedding, participants, adjacency pairs). Compare the complexity of a construction of independent paradigms which was equal to the sum of the complexity of the paradigms, as was given in Equation (1.13); and compare Equation (1.27). Moreover, each of these complexities is essentially just the logarithm of the root-mean-square of the Gaussian distribution that describes the dimension at issue. In other words, insofar as the variables were uncoupled

$$C(N, \varepsilon, p, q) = C(N, \varepsilon) + C(N, p) + C(N, q). \tag{2.8}$$

Finally, as may be seen, the complexity is maximum when the probability (or relative frequency) of embedding is 1/2, when the probability that participant 1 takes the floor is 1/2, and when the probability that a first pair-part is a question is 1/2. That is, the complexity is maximum when the probabilities of all possibilities (embed versus non-embed; participant 1 versus participant 2; question versus command) are equal. Organization, or the difference between maximum and actual complexity, therefore increases as there is any asymmetry between participant 1 and participant 2, between questions or commands, between embedding or non-embedding. This should make sense: a highly organized (and hence relatively predictable) sequence is one is which one participant (rather than another) usually takes the floor, in which the move made by that participant is usually a question (rather than a command), and in which the response is always an immediate answer (rather than some dispreferred move by way of an embedding). Imagine a master testing the knowledge of an acolyte, or a corporal barking out commands to a private. That is, social hierarchy (and any kind of institutional regimentation more generally) decreases conversational complexity. In short, Goffman's notion of ritual and system constraints (1981) could not be better named: the more constraints we place on a system, the more organized the system, the less complex the system.

To this simple model of discourse patterns we may add any number of parameters and variables. For example, we could take into account the caveat mentioned in the previous section, and thereby allow conversations to unfold at any level of embedding. We could allow the number of adjacency pairs to vary. And we could increase the degree of resolution, and allow the various parameters to couple. In the next two sections, we will treat the number of moves (N) as a variable, itself governed by a probability distribution. And we will take into consideration a cost-benefit analysis, such that complexity is maximized while length is minimized. The point is not to carry out any of these additions in exhaustive detail, but merely to show how easily additions can be made to this model, such that it may be extended to any degree of theoretical sophistication and empirical richness.

2.2 Scaling Up and Down the Analysis to Include Any Other Stochastic Variables

In all the random-walk models used so far, the number of moves (or length of the sequence), has been an independent variable, one not itself subject to a probability distribution. This has allowed us understand how the complexity of a sequence varies with the length of a sequence, usually as slowly increasing logarithmic function. In this way, the length of a sequence has been outside of statistical considerations. What we would like to do now is treat the number of moves in a sequence (N), and/or the length of a conversation more generally (when it can be treated as a single sequence), as itself governed by a probability distribution.

For example, suppose we want to calculate the complexity of conversations between telephone operators and customers. In such a scenario, we still need to know the usual information: the number of participants, and probability each of them speaks; the number of adjacency pairs, and the probability they are used; the probability that embedding occurs; and so forth. That is, we still need a means of calculating the complexity of a conversation (or sequence within a conversation) with N moves, as governed by some set of parameters and variables. In addition, we now need to know some extra information: the probability that, of all the conversations between telephone operators and customers, any given conversation has a certain number of moves, or lasts a certain length. For example, the conversations between operators and customers might follow a Gaussian distribution: say, on average, conversation are 30 moves long; and there is some reasonable bell-shaped width of conversations around this average (say, plus or minus 10 moves); but for most other lengths the probability is relatively small. Or, as will be taken up in the next section, it might be the case that, for example, in long-distance phone calls, the probability distribution follows a Boltzmann- or Zipf-like law: exponentially decreasing probability with increased length of conversation. In short, the details of the probability distribution itself are not so much at issue; what matters is that the length of conversations within some set of conversations (however bounded or defined) is governed by one.

Assuming we have a probability distribution $P(N)$, which describes the relative probability that a conversation has N moves, and assuming we know how to calculate the complexity $C(N)$ of a conversation of N moves (through the usual methods, described in previous sections), we can calculate the complexity of a set of conversations governed by these two distributions. Note, then, that what is at issue is not the complexity of a conversation *per se*, but the complexity of a field of conversations, the complexity of a genre of discourse: all calls between operators and customers; all sales at a particular store; all job interviews at a particular company; all confessions in a particular parish; and so forth. As demonstrated in the appendix, the complexity of such a field of conversations may be calculated using the following equation:

$$C = \sum_N P(N)C(N) - \sum_N P(N)\log_2 P(N). \qquad (2.9)$$

Here the summation N is over all possible lengths of conversation (say, $N=2, 4, 6, 8$, etc.); we are assuming that $P(N)$, as the probability that a conversation has a length N, is suitably normalized; and we are assuming that $C(N)$, as the complexity of a conversation of length N, has been adequately calculated. While Equation (2.9) may look strange at first, it has a very simple interpretation: the complexity of the field of conversations is just the average of the complexities (within that field) plus the complexity of the distribution (which governs that field).

Suppose, for example, that conversation length is governed by a Gaussian distribution whose average length is N_{avg}, and whose dispersion is N_{dis}. And suppose that the conversations themselves may be described by our simplest model, using $C_N(\varepsilon, p, q)$ from Equation (2.5). Using Equation (2.9), the complexity of this field of conversations is therefore given by

$$C(N_{avg}, N_{dis}) = \sum_{N=0}^{\infty} C_N(\varepsilon, p, q) \frac{2 \times \exp^{\frac{-(N-N_{avg})^2}{2N_{dis}}}}{\sqrt{2\pi N_{dis}}} - \sum_{N=0}^{\infty} \frac{2 \times \exp^{\frac{-(N-N_{avg})^2}{2N_{dis}}}}{\sqrt{2\pi N_{dis}}} \times \log_2 \frac{2 \times \exp^{\frac{-(N-N_{avg})^2}{2N_{dis}}}}{\sqrt{2\pi N_{dis}}}. \tag{2.10}$$

Here the summation only counts over even N (which explains the factor of 2, which is required for normalization); and, while this summation may include conversations of any length (up to $N=$ infinity), after a certain point the probability of a conversation with large N becomes vanishingly small – and hence the summation can be terminated at any convenient place (given the values of N_{avg} and N_{dis}). The first term on the right of Equation (2.10) is just the average complexity of a conversation within the field of conversations; and the second term on the right of Equation (2.10) is just the complexity of the probability distribution which governs that field. Crucially, such a field of conversations has a maximum complexity when the complexity of any conversation of length N is maximum (the variables ε, p, and q would all be equal to 1/2); and when the probability that a conversation has length N is proportional to the relative complexity of a conversation with length N (that is, the more complex a conversation, the more probable a conversation).

The foregoing considerations are much more general than they seem. In particular, rather than consider the length of a conversation to be governed by some probability distribution, we may consider *any parameter or variable* of the model to be governed by a probability distribution. In particular, suppose now that within some larger set of conversations, various subsets of this set occur with a certain probability – where each subset may be modelled by particular types of parameters, and particular values of variables; and suppose that some probability distribution governs the probability that a conversation within the larger set is confined to one of the subsets (and hence may be described using the types of parameters and values of variables appropriate to that subset). In short, rather than the subsets being conversations with particular lengths (as just described), the subsets can be conversations with different probabilities of embedding, different numbers or kinds of adjacency pairs (with different probabilities of usage), different numbers or types of participants (who themselves might correlate with different probabilities of embedding, or different probabilities of using adjacency pairs), and so forth. For example, such a probability distribution might govern what kinds of participants are involved in a conversation, as a function of sociological categories such as gender, age, occupation, and so forth; and such sociological categories might correlate with different values of variables. Or, such a probability distribution might govern the kind of discourse genre that unfolds: from chatting to board meetings, from lectures to scoldings – where such genres are themselves describable with different types of parameters and values of variables.

As mentioned in the caveats of Section 2.1, a particularly important type of subset has been implicitly assumed all along. In particular, the models were designed for understanding the sequencing of moves. Hence, while the preceding discussion has sometimes been couched in terms of "conversations of a particular length", we have been glossing over the ways in which conversations often consist of relatively ordered sequences – not just opening sequences (such as greetings), "guts", and closing sequences (such as goodbyes); but also all the various twists and turns that the "guts" may hold. In other words, just as we may account for the distribution of lexical and grammatical categories within moves, and just as we may account for the distribution of moves within sequences, we may also account for the distribution of sequences within conversations. In particular, different types of sequences may have different types of parameters and different values of variables. Their length may be subject to different probability distributions and the number, type, and ordering of sequences within a conversation may be analysed in terms of a random-walk model. In short, this framework allows us to scale up or down the analysis to any degree of resolution, and take into account any frame of relevance. Such iterative applicability is where this measure of complexity, and this model of conversation, gets its real power.

In each of the examples just offered, then, we can use the generalization of Equation (2.9). In particular, for any set of conversations composed of subsets, the complexity of the set is given by the average of the complexities (of the subsets), plus the complexity of the distribution (of the subsets). Formally, this may be stated as follows: anytime we know the complexity of various subsets (C^1, C^2, $C^3, \ldots, C^n$), and we know the probability (or relative frequency) that the overall set is confined to one of its subsets (such that $P_1+P_2+P_3+\ldots+P_n=1$), then the actual complexity of the overall set is given by

$$C^{1+2+3+\cdots+n} = \sum_{i=1}^{n} P_i C^i - \sum_{i=1}^{n} P_i \log_2 P_i. \qquad (2.11)$$

Equation (2.11) is just the generalization of Equation (2.9). It allows our calculations of complexity to be used iteratively – not just applicable to subsets within a set, but also to sub-subsets within a subset, and so forth – up and down to any degree of resolution, such that any and all of the foregoing issues may be considered at once.

2.3 Conclusion: Adding Economic Constraints to Models of Conversation

Perhaps more interesting than a Gaussian distribution, as modelled in Equation (2.10), is a Boltzmann- or Zipf-like distribution, as derived via microeconomic considerations. In particular, suppose that conversation length is governed by the following three considerations.

First, the larger the complexity of a conversation the better – from the standpoint of information transmission, social networking, "face-time", and so forth. That is, we would like to maximize the average complexity of a conversation.

Second, the shorter the length of a conversation the better – from the standpoint of clock-time, money, efficiency, and so forth. That is, we would like to minimize the average length of a conversation.

Third, complexity and length come with commensurable "costs" (or "benefits"), such that complexity is positively valued by some variable ($Q_C > 0$), and length is negatively valued by some variable ($Q_L < 0$). In other words, what we are really trying to do is *maximize* the function

$$U = Q_C C + Q_L \sum_N P(N)N = Q_C\left(\sum_N P(N)C(N) - \sum_N P(N)\log_2 P(N)\right) + Q_L \sum_N P(N)N. \qquad (2.12)$$

The first term on the right of the first equality in Equation (2.12) is just the (positive) value of complexity times the amount of complexity; and the second term on the right of the first equality in Equation (2.12) is just the (negative) value of length times the average length. The first term on the right of the second equality in Equation (2.12) merely incorporates Equation (2.9). $P(N)$ is subject to the usual normalization constraint, such that

$$\sum_N P(N) = 1. \qquad (2.13)$$

The complexity of a conversation of a given length, $C(N)$, may be calculated however one chooses (given the types of conversations one is studying). For present purposes, we will use Equation (2.7), which follows from the Gaussian approximation to the simplest conversational patterns. As derived in Appendix B, the probability distribution $P(N)$ which satisfies these requirements is

$$P(N) = K\sqrt{(2\pi e(N/2-1)\varepsilon(1-\varepsilon))(2\pi e(N/2)p(1-p))(2\pi e(N/2)q(1-q))} \times 2^{Q_L/Q_C N}. \qquad (2.14)$$

The first term in this equation is just a constant, which is to be determined by the normalization constraint given in Equation (2.13). As may be seen by comparing Equation (2.14) with Equation (2.7), the second term is just the product of the three root-mean-square deviations underlying the simplest conversational patterns. This term is a slowly increasing function of N. The third term is the important term. Given the fact that Q_L is negative and Q_C is positive, this describes a rapidly decreasing function of N. Indeed, this is an exponentially decreasing function of N, and hence quickly cancels out the second factor. In short, for small N, $P(N)$ is an increasing function (with the second term dominating); and then for large N, $P(N)$ is a rapidly decreasing function (with the third term dominating). As may be seen in Figure 2.3, the degree to which this domination takes place is a function of the ratio between Q_L and Q_C. When this ratio is -0.2, the average length of a conversation is about 19 moves, and the complexity is 4.30 bits; when this ratio is -0.1, the average length of a conversation is about 37 moves, and the complexity is 5.33 bits; and when this ratio is -0.05, the average length of a conversation is about 73 moves, and the complexity is 6.34 bits. In short, the more complexity is (positively) valued, the less rapidly the function decreases (such that conversations with longer lengths are more probable) and the more length is (negatively) valued, the more rapidly the function decreases (such that conversations with longer lengths are less and less probable).

While Equation (2.14) was exemplified using the model of conversational complexity given by Equation (2.7), it is much more general than that. In particular, given any model of conversational complexity which

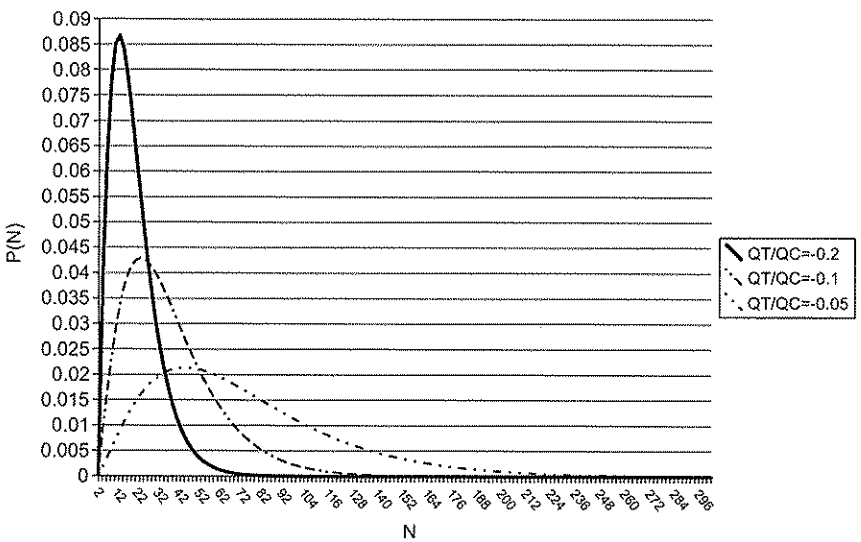

Fig. 2.3. Maximizing complexity and minimizing length.

provides some function $C(N)$, and given similar cost-benefit considerations, the probability distribution which satisfies all the above constraints may be represented as

$$P(N) = K \times 2^{C(N)} \times 2^{Q_L/Q_C N}. \qquad (2.15)$$

What is so special about Equation (2.15) is not the details of the probability distribution *per se*; rather, it is the fact that we have theoretically derived a probability distribution (rather than having empirically observed it, as all the preceding analysis has presupposed). This theoretical derivation turns on the calculus of variations (the maximization or minimization of certain functions) as applied to microeconomic considerations (the costs and benefits of various values of variables and/or types of parameters). The costs and benefits, needless to say, need not be economic (such as those constraining how long the average cell-phone conversation is, depending on the typical provider and plan); but may, rather, turn on cultural values (politeness norms, status hierarchies, etc.), cognitive processes (such as limits on attention, memory, information-transmission, etc.), and so forth. Here, then, is a key locale where *value* (in the widest sense) enters the model, thereby providing a central constraint on, and condition for, the meaningful organization of conversational complexity.

REFERENCES

Brillouin, L. (1962). *Science and Information Theory*, 2nd edition. New York: Academic Publishers.

Boltzmann, L. (1877). Über die Beziehung eines allgemeinen mechanischen Satzes zum zweiten Hauptsatz der Warmtheorie. Sitzungsbericht der Akademie der Wissenschaften Wien. *Math.-Naturwiss. Kl.*, *75*, 67–73.

Brooks, D. R., & Wiley, E. O. (1988). *Evolution as Entropy*, 2nd edition. Chicago: University of Chicago Press.

Duranti, A. (1997). Polyphonic discourse: Overlapping in Samoan ceremonial greetings. *Text*, *17*, 349–81.

Givón, T. (1983). *Topic Continuity in Discourse*. Amsterdam and Philadelphia: John Benjamins.

Goffman, E. (1991). Replies and responses. In *Forms of Talk* (pp. 5–76). Philadelphia: University of Pennsylvania Press.

Greenberg, J. (1966). *Language Universals*. The Hague: Mouton & Co.

Jefferson, G. (1984). Notes on some orderliness of overlap onset. In V. D'Urso & P. Leonardi (Eds), *Discourse Analysis and Natural Rhetorics* (pp. 11–38). Padova: CLEUP Editore.

Kauffman, S. A. (1993). *The Origins of Order: Self-Organization and Selection in Evolution*. New York: Oxford University Press.
Layzer, D. (1988). Growth of order in the universe. In B. H. Weber, D. J. Depew & J. D. Smith (Eds), *Entropy, Information, and Evolution: New Perspectives on Physical and Biological Evolution* (pp. 23–40). Cambridge: MIT Press.
Ochs, E., Schegloff, E., & Sandra Thompson (1996). *Interaction and Grammar*. Cambridge: Cambridge University Press.
Reif, F. (1965). *Fundamentals of Statistical and Thermal Physics*. Boston, MA: McGraw Hill.
Sacks, H., Schegloff, E., & Gail, J. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*, 696–735.
Schegloff, E. (1986). The routine as achievement. *Human Studies*, *9*, 111–151.
Schegloff, E. (1996). Turn organization: One intersection of grammar and interaction. In E. Ochs, E. A. Schegloff, & S. A. Thompson (Eds), *Interaction and Grammar* (pp. 52–133). Cambridge: Cambridge University Press.
Schegloff, E. (2004). Interaction. Paper presented at Wenner-Gren Foundation Symposium on the Roots of Human Sociality: Culture, Cognition, and Human Interaction, Duck, New Hampshire, October.
Schegloff, E. (2005). *A Primer in Conversational Analysis: Sequence Organization*. Cambridge: Cambridge University Press.
Schegloff, E., & Sacks, H. (1973). Opening up closings. *Semiotica*, *8*, 289–327.
Schegloff, E., Gail, J., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*(2), 361–382.
Shannon, C. E. (1963 [1948]). *The Mathematical Theory of Communication*. Urbana and Chicago: University of Illinois Press.
Simon, H. A. (1996). *The Sciences of the Artificial*, 3rd edition. Cambridge: MIT Press.
Tolman, R. C. (1979 [1938]). *The Principles of Statistical Mechanics*. New York: Dover.
Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge: Harvard University Press.

APPENDIX A: THE DERIVATION OF EQUATION (2.9) AND EQUATION (2.11)

Suppose a system is composed of n subsystems, each of which has a complexity given by

$$\begin{aligned} C^1 &= -\sum_i P_i^1 \log_2 P_i^1 \\ C^2 &= -\sum_i P_i^2 \log_2 P_i^2 \\ &\vdots \\ C^n &= -\sum_i P_i^n \log_2 P_i^n \end{aligned} \tag{A.1}$$

Here the summation for each subsystem is over all possible states of that subsystem, howsoever many there may be, and the probability distribution of each subsystem is suitably normalized, such that

$$\sum_i P_i^1 = \sum_i P_i^2 = \cdots = \sum_i P_i^n = 1. \tag{A.2}$$

Suppose that the probability that the system is found in one of its n subsystems is given by

$$P = (P_1,\ P_2,\ \cdots,\ P_n) \tag{A.3}$$

which is suitably normalized, such that

$$\sum_{j=1}^{n} P_j = 1. \tag{A.4}$$

Then the complexity of the total system is just

$$C^{1+2+\cdots+n} = -\sum_{j=1}^{n}\sum_i P_j P_i^j \log_2 P_j P_i^j. \tag{A.5}$$

This reduces to

$$C^{1+2+\cdots+n} = -\sum_{j=1}^{n}\sum_i P_j P_i^j \log_2 P_j - \sum_{j=1}^{n}\sum_i P_j P_i^j \log_2 P_i^j. \tag{A.6}$$

Given the normalization constraints, Equation (A.3) and Equation (A.4), this reduces to

$$C^{1+2+\cdots+n} = -\sum_{j=1}^{n} P_j C^j - \sum_{j=1}^{n} P_j \log_2 P_j, \tag{A.7}$$

and Equation (A.7) is equivalent to Equation (2.11). Finally, by treating P_j as $P(N)$ and C^j as $C(N)$, one recovers Equation (2.9).

APPENDIX B: THE DERIVATION OF EQUATION (2.14) AND EQUATION (2.15)

Let Q_C be constant which denotes the positive value of complexity, such that $Q_C > 0$. Let Q_L be a constant which denotes the negative value of length, such that $Q_L < 0$. Let P_N be the probability that a conversation has length N. And let C_N be the complexity of a conversation of length N, taking into account whatever parameters and variables are relevant. Following Equation (2.12), we seek to maximize the function

$$\begin{aligned} U &= Q_C C + Q_L \sum_N P_N N \\ &= Q_C \left(\sum_N P_N C_N - \sum_N P_N \log_2 P_N \right) + Q_L \sum_N P_N N \end{aligned} \tag{B.1}$$

subject to the constraint that

$$\sum_N P_N = 1. \tag{B.2}$$

Using the method of Lagrange multipliers, the Lagrangian is given by

$$L = Q_C \left(\sum_N P_N C_N - \sum_N P_N \log_2 P_N \right) + Q_L \sum_N P_N N + \alpha \left(\sum_N P_N - 1 \right) \tag{B.3}$$

(We are assuming that C_N is not a function of P_N.) Lagrange's theorem says that an optimal choice of $P = \{P_1, P_2, \ldots, P_N\}$ must satisfy the $N+1$ first-order conditions

$$\begin{aligned} \frac{\partial L}{\partial P_1} &= Q_C \left(C_N - \log_2 P_1 - \frac{1}{\ln 2} \right) + Q_L N + \alpha = 0 \\ \frac{\partial L}{\partial P_2} &= Q_C \left(C_N - \log_2 P_2 - \frac{1}{\ln 2} \right) + Q_L N + \alpha = 0 \\ &\vdots \\ \frac{\partial L}{\partial P_N} &= Q_C \left(C_N - \log_2 P_N - \frac{1}{\ln 2} \right) + Q_L N + \alpha = 0 \\ \frac{\partial L}{\partial \alpha} &= \sum_N P_N - 1 = 0 \end{aligned} \tag{B.4}$$

The last of these is just the initial constraint, Equation (B.2). And each of the others reduces to

$$\log_2 P_i = \frac{Q_L}{Q_C} N + \frac{\alpha}{Q_C} + C_i - \frac{1}{\ln 2}, \quad \textit{for } i = 1, 2, \cdots, N \tag{B.5}$$

This reduces to

$$P_i = 2^{\frac{Q_L}{Q_C} N} \times 2^{C_i} \times 2^{\frac{\alpha}{Q_C}} \times 2^{\frac{-1}{\ln 2}}, \quad \textit{for } i = 1, 2, \cdots, N \tag{B.6}$$

Setting the second two terms equal to the constant K, whose value is to be determined by the normalization constraint, Equation (B.2), this is equal to

$$P_i = K \times 2^{\frac{Q_L}{Q_C} N}, \quad \textit{for } i = 1, 2, \cdots, N \tag{B.7}$$

Letting $P_i = P(N)$ and $C_i = C(N)$, letting N range from 1 to any number, this is equivalent to Equation (2.15), and taking $C(N)$ from Equation (2.7), this reduces to Equation (2.14).