Quantifying and Measuring Morphological Complexity

Max Bane
University of Chicago

1 Introduction[1]

It is a standard assumption in Linguistics that all human languages are equally (and enormously) complex; when looked at as a whole, no language can be called "simpler" than another. Certainly, languages can differ in the distribution of their complexity, so that one might employ a richer inflectional system, or entertain a more complicated gamut of syllable shapes than another, but it is generally supposed that these differences must "even out" as one considers entire linguistic systems. Where there is atypical simplicity in morphology, for instance, it is assumed that one will find compensatory complexity in possible syntactic distinctions, or subtle lexical differences, or something else.

A number of researchers have recently begun to approach this equal complexity hypothesis as an empirical claim to be tested under particular definitions of complexity. Perhaps the most famous recent example is McWhorter's (2001) controversial claim that "creole grammars are the world's simplest grammars," but see also Juola 1998, Shosted 2006, Nichols 2007, and Pellegrino *et al.* 2007. The hypothesis deserves formal articulation and scrutiny because it suggests some important consequences, if confirmed:

i. Language change must be subject to a sort of "Newton's Third Law of Motion," whereby linguistic complexity remains constant over time. Any complexity metric that proves consistent with the equal complexity hypothesis could potentially bolster the predictive power of historical linguistics.

ii. Systems of grammatical representation might fruitfully be made subject to the constraint that no grammar of a possible human language be more complex than any other, *in toto*.

Even should we adopt a weaker version of the equal complexity hypothesis, supposing, say, that there is some narrow range of complexity in which all human languages fall, it remains an important task to map the boundaries of that range according to some metric.

In this paper I argue for an information theoretic approach to defining linguistic complexity, and offer preliminary results for a novel method of applying this approach to natural language morphology. Section 2 gives a brief review of some existing complexity metrics and proposes one for morphology based on the information theoretic notion of Kolmogorov complexity. Section 3 describes the results of an exploratory application of this metric to twenty languages, and section 4 then discusses work to be done and the problem of evaluating complexity metrics in general.

2 Defining Complexity

2.1 Existing Metrics

There is of course no generally agreed on definition of grammatical complexity in linguistics, though a desire for one can be traced back to the evaluation metrics of early generative grammar, and even further to Goodman (1943). A common method of quantifying complexity in recent work (McWhorter 2001, Bickel and Nichols 2005, Shosted 2006, Nichols 2007) is to count the occurrences of a variety of hand-picked, intuitively justified properties, or complexity "indicators," as Shosted (2006:3) calls them, of the linguistic system in question. For example, one might relate phonological complexity to the size of a language's phoneme and syllable inventories, to the number of "marked" forms it permits, or to

[1]I would like to thank Adam Albright, John Goldsmith, Patrick Juola, Daniel Karvonen, Jason Merchant, Salikoko Mufwene, Jason Riggle, and Alan Yu for valuable feedback and stimulating discussion relating to this work. Special thanks are due to Kevin Scannell for his assistance in obtaining a number of corpora.

Cascadilla Proceedings Project
Completed June 30, 2007

the number of phonological alternations known to occur. In the case of morphological complexity, the number of possible inflection points in a "typical" sentence, the number of inflectional categories, and the number of morpheme types are all likely candidates for complexity indicators. A metric of syntactic complexity could conceivably be gotten from the number of parameters of variation deviating from their presumed defaults values, or from the number of syntactic operations that must apply in an average sentence.

Though this general approach to quantifying complexity is often useful, it does suffer from some drawbacks. Chief among them is the question of how to decide in a principled way which of the myriad linguistic properties that can be measured should be admitted into the set of complexity indicators. Additionally, even supposing one has a good guiding principle for the selection of complexity indicators, constructing a unified metric as a function of these indicators often requires answering questions that do not have intuitively obvious answers: how many marked syllables should a given phonological alternation be worth in terms of complexity?; and so on.

An alternative approach is to attempt to apply existing, independently motivated conceptions of complexity, as studied in information theory, to linguistic systems. Information theory offers at least two such formalisms that might be appropriated. The first is information entropy (Shannon 1948), a quantity defined with respect to a probability model, that captures the average number of bits of information necessary to specify the state of a random variable or system described by that probability model. Pellegrino *et al.* (2007), del Prado Martín *et al.* (2004) and others have used this notion to measure linguistic complexity, with interesting results. The second possibility is Kolmogorov complexity (Solomonoff 1964, Kolmogorov 1965, Chaitin 1987), a more general quantity that gives an absolute and objective measure of the information content of an object, independent of any probability model. Juola (1998, in press) has experimented with a text-deformation method of measuring linguistic complexity based on this formalization; the metric I propose in this paper is also grounded in Kolmogorov complexity (and indeed replicates some aspects of Juola's results; see section 3.2), though it employs a very different method.

Neither entropy nor Kolmogorov complexity are without limitations if one wishes to apply them to aspects of language. The former requires that the atomic units of the system whose complexity one is measuring be statistically independent of each other, a constraint that is difficult to satisfy in linguistic systems (Pellegrino *et al.* (2007) give a good discussion of the issues involved, and in their case settle on the syllable as an approximately independent unit), while the latter is not directly calculable, forcing one to depend on computable approximations of Kolmogorov complexity.

2.2 Kolmogorov Complexity

McWhorter (2001), in describing the intent of his proposed complexity metric, hits upon the intuition at the heart of Kolmogorov complexity: "My object of inquiry is differentials between grammars in degree of overspecification. . . to the extent that some grammars might be seen to require *lengthier descriptions in order to characterize even the basics of their grammar than others.*" (pp. 134–5, emphasis mine). This idea that one object is more complex than another insofar as it takes longer to describe is precisely what Kolmogorov complexity aims to formalize, as a property of strings. Namely, it is the length of the shortest description, in some agreed on description language L, of the string in question. Consider two strings over the alphabet $\{0, 1\}$:

(1) a. 10101010101010101010
    b. 11011111000101011010

String (1a) can be described in English as "*10* ten times," while (1b) has no obvious such description in English, beyond listing the string itself. If we take English as our L, and if we measure the length of a description in English as the number of characters it comprises in standard orthography, then it would seem, tentatively, that (1a) is the simpler of the two. Crucially, though, to make such a claim within the framework of Kolmogorov complexity, one would have to demonstrate that there is no shorter description of (1a) in the chosen description language, and that there is no shorter description of (1b) than simply listing it. One might be surprised to learn that (1b) can in fact be described rather succinctly in English as "913754 in binary," or if the listener is assumed to possess some relevant background knowledge, perhaps simply as "913754 BIN," which is a shorter description than that given for (1a).

Language	*Bits*	*Language*	*Bits*
Danish	7,159,576	Dutch	8,182,264
English	6,895,608	French	8,240,232
German	8,039,792	Haitian Creole	7,298,360
Hungarian	8,163,704	Icelandic	7,953,120
Italian	9,049,912	Latin	7,887,288
Maori	7,064,968	Spanish	7,412,232
Swedish	7,597,400		

Table 1: Upper bounds on the Kolmogorov complexity of the Bible, as translated into thirteen languages; computed by the `bzip2` compression algorithm.

This illustrates two important points. The first is that English is not a very good choice for a description language; there is simply no rigorous definition of what it means for an English string to "describe" something, hence a certain indeterminacy as to whether something like "913754 BIN" should be considered a valid description, contingent upon shared knowledge, etc. The other is that, even supposing one has chosen a useful description language, it might prove very difficult to ever be sure that one has truly found the shortest possible description of a given string.

The first concern, that the description language carry a well defined sense of what it means to describe, can be addressed by choosing a computer language as L. That is, a description of a string S is the code C (in a chosen programming language L) of any program that outputs S. If C represents the smallest possible such program, then the Kolmogorov complexity of S is $|C|$. Theorists usually avoid wedding themselves to any particular programming language by instead using a standard encoding for the Gödel numbers of the universal Turing machine, though the details of this theoretical convenience are not important for what follows. One important point is that we can always assume without loss of generality that both the string to be described, S, and the code that describes it, C, are expressed in the binary alaphabet $\{0, 1\}$; thus the units of Kolmogorov complexity are bits, i.e., units of information.

The second concern, that for a given string, we might never be sure we've found its shortest description in the description language, is in fact confirmed by one of the fundamental results of algorithmic information theory: even though the Kolmogorov complexity of a string is a well defined, objective quantity, one can never actually compute it directly (see, e.g., Chaitin 1987 for a proof). In practice, one must instead employ approximation methods that establish upper bounds on Kolmogorov complexity. Juola (1998), for instance, depends on the standard `gzip` compression algorithm as one such approximation; another possibility is described in the next section.

Thus, to construct a complexity metric for some domain on the basis of Kolmogorov complexity, one must answer three questions: (i) what exactly is the object or system whose complexity is to be measured? (ii) how shall that object be encoded as a (binary) string? and (iii) what method shall be used to approximate the Kolmogorov complexity of that string? Suppose, for a simple example, that we are interested in a complexity metric for the expression, in a given human language λ, of the Bible, suitably translated. A convenient string encoding of such an object is simply the text of the Bible in λ, according to λ's standard orthographic conventions (this encoding is easily reduced to a binary one). The Kolmogorov complexity of the resulting string might then be approximated through a standard compression algorithm. The results of such an experiment for thirteen languages are given in Table 1 (see section 3.1 for more information on the data used).

Of course, the complexity of some *corpus* in a given language is not necessarily a linguistically interesting quantity. Linguistic complexity is better conceived of as the complexity of a language's *grammar*, or some component thereof (the phonological grammar, morphological grammar, etc.). The situation is schematized in (2).

(2) $D \rightarrow G \rightarrow \lambda$

G is a formal grammar devised by linguists to generate, or account for, the observed properties of some aspect (phonological/morphological/syntactic/...) of a linguistic system, λ. If D is the shortest description of (i.e., computer program that outputs) a suitable string encoding of G, then $|D|$ is the Kolmogorov

complexity of the grammar for λ. In practice, D will always be an approximation of the shortest description, so that $|D|$ is an upper bound on the Kolmogorov complexity. The job of constructing a complexity metric for λ, then, consists of deciding what form its grammar G will take, and which approximation method to use.

2.3 Minimum Description Length and Morphology

Morphology is a good domain in which to begin experimenting with proposed complexity metrics for several reasons. As McWhorter (p.c. in Shosted 2006) notes, it displays a "usually richer and more widespread interaction with syntax, this interaction being of note in covering the general issue of complexity more widely." Juola (1998:2) remarks that morphology is an obvious "testbed" for theories of measuring complexity, since "it is intuitively apparent that some languages (for example, Finnish) are 'morphologically complex' while others are more simple. On the other hand, claims about (e.g.) semantic differences [in complexity] are less intuitive and less widely accepted."

Morphology is also a convenient place to begin applying information theoretic complexity metrics beacuse of recent work by Goldsmith (2001, 2006) and Hu (2007) on the application of general complexity-based inductive methods to the automatic learning of natural language morphology; their results have been collected into a freely available software package called Linguistica.[2] The intent of Linguistica is not to measure linguistic complexity *per se*, but rather to create a computer program that emulates what a linguist does when analyzing the morphology of a language, solely on the basis of example texts. It does so by application of a general inductive technique based on "minimum description length," (MDL; Rissanen 1984) a method of approximating Kolmogorov complexity that is rendered computable by restricting the class of allowed descriptions (though in MDL theory, one usually speaks of candidate models or hypotheses rather than descriptions). The essential method followed by Linguistica is to attempt to construct as small a model of the data (in the allowed class) as possible, that simultaneously predicts the data as well as possible.

At the risk of simplifying for our present purposes (see Goldsmith 2001 for more details), the models (descriptions) that Linguistica constructs are lexica consisting of stems, affixes, and what Goldsmith calls "signatures," which describe the possible distributions of affixes upon stems. Here are some example entries in a lexicon induced by Linguistica from a corpus of Standard French:

(3)

	Stem	*Suffixal Signature*
a.	accompli	∅.e.t.r.s.ssent.ssez
b.	académi	cien.e.es.que
c.	académicien	∅.s

Entry (3a) indicates that the stem *accompli-* can take the suffixes -∅ (masculine past participle), *-e* (feminine past participle), *-t* (third person singular), *-r* (infinitive), *-s* (second person singular), *-ssent* (third person plural), *-ssez* (second person plural). Thus the signature ∅.e.t.r.s.ssent.ssez corresponds to something like the inflectional category "verbs in *-ir*" in French. Similarly, (3b) indicates that *académi-* is the stem of words like *académicien* "academician," *académie* "academy," *académies* "academies," *académique* "academic." Furthermore, *académicien* is itself a stem which can take singular -∅ and plural *-s*, as shown in (3c).

Linguistica reads in a corpus of text in the target language and iteratively applies a series of heuristics to find the simplest model (i.e., lexicon as in (3)) that best describes the corpus. That is, for each stem, affix, and signature a description length is calculated and tracked, and the "simplest" model in this case is that with the smallest total description length over all stems, affixes, and signatures.[3] These description lengths are approximations, or indices, of complexity (in the Kolmogorov sense), so that a lexicon's total description length is an approximation of its complexity.

[2] Available at http://linguistica.uchicago.edu.

[3] More properly, Linguistica attempts to minimize the complexity (description length) of the model *summed with* the complexity of the corpus given the model as a probabilistic generator of it.

2.4 A Metric

How then might we go from the complexity of the lexicon generated for a particular language to the complexity of that language's morphology? The total complexity of a lexicon is distributed between (i) its stems, (ii) its affixes, and (iii) its signatures. For a language with few and simple inflections, most of the total information (hence, complexity) encoded by the lexicon will reside in the list of stems, whereas for a morphologically more complex language, Linguistica will apportion more of that information to the lists of affixes and signatures. Thus we can measure a language's morphological complexity as the proportion of the lexicon's total description length that is due to the description lengths of the affixes and signatures. That is, if $DL(x)$ is the description length of x,

(4) Morphological complexity $= \frac{DL(\text{Affixes})+DL(\text{Signatures})}{DL(\text{Affixes})+DL(\text{Signatures})+DL(\text{Stems})}$

In principle, we might also take morphological complexity to be the sum description length of affixes and signatures (the numerator in (4)) — it would then be a directly expressed upper bound on the Kolmogorov complexity of the relevant parts of the lexicon, expressed in bits. The reason for expressing it as a unitless ratio of description lengths as in (4) is to insulate the metric from the incidental deficiencies of available corpora, which will not generally exhibit the full range of morphological combinations possible in a language.

Such a metric is not without its limitations, of course. The greatest of these is that it relies entirely on Linguistica's inductive technique for its notion of what morphology is. Linguistica is currently poorly suited to analyzing languages that make heavy use of infixal, templatic, or reduplicative morphology, and so the proposed metric will not generally be applicable to such languages. Even in the case of mostly prefixing and suffixing languages, Linguistica's automatically discovered analyses will not always correspond to what a trained linguist might conclude. Ideally, we would cut Linguistica out of the picture entirely, and compute the metric on the basis of morphological lexica created by hand in consultation with native speakers, but this would obviously severely limit the speed at which one could survey large numbers of languages.

3 Preliminary Measurements

3.1 Corpora

To begin exploring the empirical behavior of the metric proposed above, a total of twenty languages were selected for preliminary measurements. Since Linguistica must derive its morphological analyses from samples of text, the choice of languages was largely determined by the size and quality of available corpora. Additionally, an effort was made to include a number of creole languages (seven total), in order to test whether the proposed metric reflects the often (intuitively) claimed relative simplicity of their morphology.

Fourteen of the languages are represented by corpora consisting of translations of the Bible, as made freely available on the internet.[4] These languages are listed in (5) together with the word-token counts of their copora.

(5) Danish (653,036), Dutch (727,489), English (737,241), French (728,191), German (729,853), Haitian Creole (920,332), Hungarian (597,084), Icelandic (667,363), Italian (774,946), Latin (604,305), Maori (977,565), Spanish (664,108), Swedish (675,315), Vietnamese (837,733).

The remaining six languages are creoles or pidgins for which biblical translations are not readily available, and whose corpora were gathered from the web by Scannell's (2007) automatic, language-targeting web-crawler. They are listed in (6) with token counts.

(6) Bislama (62,451), Kituba (45,275), Nigerian Pidgin (281,639), Papiementu (730,189), Solomon Pijin (19,986), Tok Pisin (1,027,044).

Each corpus was normalized for capitalization, extraneous punctuation, etc.

[4] Available at `http://www.biblegateway.com/versions/` as of June, 2007.

Language	*Metric*	*Language*	*Metric*
Latin	35.51%	English	16.88%
Hungarian	33.98%	Maori	13.62%
Italian	28.34%	Papiementu	10.16%
Spanish	27.50%	Nigerian Pidgin	9.80%
Icelandic	26.54%	Tok Pisin	8.93 %
French	23.05%	Bislama	5.38%
Danish	22.86%	Kituba	3.40%
Swedish	21.85%	Solomon Pijin	2.91%
German	20.40%	Haitian Creole	2.58%
Dutch	19.58%	Vietnamese	0.05%

Table 2: Computed values of the proposed ratio metric (4) for all twenty languages surveyed.

3.2 Results

From each corpus, Linguistica induced a morphological lexicon of stems, prefixes, suffixes, and signatures describing their possible combinations. These were output together with their computed description lengths, yielding sufficient information to calculate the ratio in (4). Table 2 gives the results for all corpora. One can see that all of the creoles and pidgins receive comparatively low values of morphological complexity according to the metric, together with Vietnamese. The remaining languages appear to be ranked plausibly, with Latin and Hungarian topping the list as most complex. Additionally, the relative ranking of Maori, English, Dutch, and French — the four languages shared by my survey and that of Juola (1998) — agrees with that yielded by Juola's metric.

Table 3 gives the complexity values for the fourteen languages represented by Bible corpora, together with the type and token counts of the corpora. There is an apparent trend: as the complexity metric increases, so does the type count, while the token count decreases. That is, languages that score highly according to the metric employ many word types, but fewer tokens thereof, while lower scoring languages use more tokens of fewer types. These relations are highly significant according to Spearman's rank test (Metric-Types: $\rho = 0.96$, $p < 2.2 \times 10^{-16}$; Metric-Tokens: $\rho = -0.76$, $p = 0.002$) and Kendall's T-test (Metric-Types: $\tau = 0.89$, $p = 1.81 \times 10^{-7}$; Metric-Tokens: $\tau = -0.60$, $p = 0.002$).

Juola (1998) finds a similar relation between his text-deformational metric of morphological complexity and the type and token counts of his corpora. This is a plausible result that one might expect to see if the metric is indeed measuring something like morphological complexity, given that the corpora involved are all purported translations of each other. Distinctions or constructions that are expressed in languages like Latin or Hungarian by morphologically altering the shapes of words (each different shape corresponding to a different word-type), are often expressed in morphologically simpler languages by the presence of function words or particular syntactic arrangements, each instance of which corresponds to new tokens of already existing types. Hence a rising token count as morphological complexity decreases, and a rising type count as morphological complexity increases.

4 Discussion

4.1 Evaluating Complexity Metrics

The approach taken in this paper has been to begin with an *a priori*, mathematical notion of complexity, which is essentially an elaborated statement of our intuitive grasp of the difference between "complex" and "simple" objects in the most general sense ("an object is more complex insofar as it takes longer to describe"), and to examine how it might be applied to linguistic systems like the morphological component of a grammar. A related, but separate question that I have not considered so far is: given two competing, proposed metrics of linguistic complexity, on what basis does one choose between them? How should we evaluate our complexity metrics?

Language	*Metric*	*Types*	*Tokens*
Latin	35.51%	46,722	604,305
Hungarian	33.98%	63,046	597,084
Italian	28.34%	35,232	774,946
Spanish	27.50%	29,021	664,108
Icelandic	26.54%	34,911	667,363
French	23.05%	31,684	728,191
Danish	22.86%	24,280	653,036
Swedish	21.85%	23,964	675,315
German	20.40%	24,692	729,853
Dutch	19.58%	21,242	727,489
English	16.88%	15,570	737,241
Maori	13.62%	8,271	977,565
Haitian Creole	2.58%	7,307	920,332
Vietnamese	0.05%	7,144	837,733

Table 3: Computed metric values and type/token counts for the fourteen biblical corpora.

This is a difficult question to answer without some objective notion of what properties "complex" linguistic systems should have, beyond simply requiring lengthier descriptions. One possibility is to adopt a processing perspective, and demand that a proposed complexity metric correlate with the sorts of things one might reasonably expect it to under a particular processing model; e.g., rates of speech errors and disfluencies, childhood and L2 acquisition properties, etc. Rigorously elucidating the expected effects of complexity according to a given metric in a particular processing model may not always prove easy, though. A second possible stance is essentially pragmatic: in a sense, it doesn't really matter whether a metric truly corresponds to whatever we mean by "complexity," as long as it is useful; for example, if one finds a metric that empirically turns out to be invariant across languages, that might prove to be a very powerful predictive tool, regardless of whether it's actually "complexity," and would furthermore motivate much research into the reasons for its invariance.

The question will have to be left open for now.

4.2 Future Work

The method outlined here for basing a metric of morphological complexity on the information theoretic notion of Kolmogorov complexity is only one possibility. Juola (1998, in press) offers an alternative, somewhat simpler, method of applying Kolmogorov complexity to morphology. His approach is to deform a corpus in such a way as to efface any morphological information from the text, and to approximate the Kolmogorov complexity of the corpus before and after this effacement, so that the ratio of these "before" and "after" measurements gives an indication of the language's morphological complexity. The effacement is accomplished by replacing each word-type with a random integer, thus obscuring any original linguistic information internal to the word, while retaining external (presumably syntactic) information about the ordering and collocations of the word's tokens. One problem with this method is that the phonological content of each word is effaced along with any morphological information, so that one ends up measuring, in effect, some mixture of phonological and morphological complexity. The Linguistica-based method described in this paper also conflates phonolgical and morphological complexity to a certain extent, since the phonological content of each affix counts towards its description length; however, since the actual morphemes have been identified, it should be possible to separate the purely phonological information from that needed to track the existence and distribution of each morpheme. This indicates a possible line of future research into refining the metric.

Additional directions for future work might include investigating the potential for correlations between proposed complexity metrics and the kinds of processing phenomena mentioned above (speech error rates, etc.), and experimenting with how metrics based on Kolmogorov complexity might be extended to grammatical domains other than morphology, with an eye toward establishing an information

theoretic typology of linguistic complexity.

References

BICKEL, BALTHASAR, and JOHANNA NICHOLS. 2005. Inflectional synthesis of the verb. In *The World Atlas of Language Structures*, ed. by Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie, 94–97. Oxford: Oxford University Press.

CHAITIN, J. G. 1987. *Algorithmic Information Theory*. Cambridge: Cambridge University Press.

DEL PRADO MARTÍN, MOSCOSO, ALEKSANDAR KOSTIC, and R. HARALD BAAYEN. 2004. Putting the bits together: An information-theoretical perspective on morphological processing. *Cognition* 94.1–18.

GOLDSMITH, JOHN. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27.153–98.

——. 2006. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering* 12.1–19.

GOODMAN, NELSON. 1943. On the simplicity of ideas. *Journal of Symbolic Logic* 8.107–21.

HU, YU, 2007. *Topics in Unsupervised Language Learning*. University of Chicago dissertation.

JUOLA, PATRICK. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics* 5.206–13.

——. in press. Assessing linguistic complexity. In *Typology, Contact, Change*, ed. by Fred Karlsson. Amsterdam: John Benjamins Press.

KOLMOGOROV, A. N. 1965. Three approaches to the quantitative defintion of information. *Problems in Information Transmission* 1.1–7.

MCWHORTER, JOHN. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5.125–66.

NICHOLS, JOHANNA. 2007. The distribution of complexity in the world's languages. In *81st Annual Meeting of the Linguistic Society of America*.

PELLEGRINO, F., C. COUPÉ, and E. MARSICO. 2007. An information theory-based approach to the balance of complexity between phonetics, phonology, and morphosyntax. In *81st Annual Meeting of the Linguistic Society of America*.

RISSANEN, J. 1984. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory* IT-30.629–36.

SCANNELL, KEVIN P., 2007. Corpus building for minority languages. http://borel.slu.edu/crubadan, accessed June, 2007.

SHANNON, CLAUDE E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27.379–423, 623–656.

SHOSTED, RYAN. 2006. Correlating complexity: A typological approach. *Linguistic Typology* 10.1–40.

SOLOMONOFF, R. J. 1964. A formal theory of inductive inference. *Information and Control* 7.1–22, 224–54.