Assessing Linguistic Complexity

Patrick Juola

Duquesne University

juola@mathcs.duq.edu

Abstract

The question of "linguistic complexity" is interesting and fruitful. Unfortunately, the intuitive meaning of "complexity" is not amenable to formal analysis. This paper discusses some proposed definitions and shows how complexity can be assessed in various frameworks. The results show that, as expected, languages are all about equally "complex," but further that languages can and do differ reliably in their morphological and syntactic complexities along an intuitive continuum.

I focus not only on the mathematical aspects of complexitgy, but on the psychological ones. Any claim about "complexity" is inherently

about process, including an implicit description of the underlying cognitive machinery. By comparing different measures, one may better understand on human language processing and similarly, understanding psycholinguistics may drive better measures.

1 Introduction

Most people with any background in language have at least an informal understanding of language complexity — a language is "complex" to the extent that you have to study in order to pass the exam on it, and in particular to the amount of stuff you simply have to memorize, such as lists of irregular verbs, case systems and declension patterns, and apparently arbitrary aspects of words such as gender. But to compare languages objectively requires a more formal specification of complexity, ideally one suited to a unified numerical measurement. Many ad-hoc complexity measures have been proposed, of which (Nichols 1986) is an obvious example; she counts the number of points in "a typical sentence" that are capable of receiving inflection.

McWhorter's definition (McWhorter 2001; McWhorter 2005) encompasses a number of similar ad-hoc measures (e.g., a language is more complex if it has more marked members in its phonemic inventory, or if it makes more

extensive use of inflectional morphology), but he ties this, at least in theory, to a single numerical measure – the length of the grammar that a language requires. Despite the obvious practical difficulties (how do you compare two different inflectional paradigms, or how do you balance simple morphology with complex phonology), this provides a reasonable formulation for judging complexity.

However, as will be discussed in the remainder of this chapter, the question of "length" itself raises the issue of in what language (in a mathematical sense, i.e. set of primitive operations) the description should be made. It is argued here that information theory provides several different approaches that yield different answers to questions of "linguistic complexity" — and that analysis of data from natural languages can shed light on the psychological and cognitive processes appropriate to such description.

2 Information Theory Basics

2.1 Zipf and Shannon

The clearest statement of the motivation of the current work on an information-theoretic basis can be found in (Zipf 1949), in his argument about applica-

tions of words:

> Man talks in order to get something. Hence man's speech may be likened to a set of tools that are engaged in achieving objectives. True, we do not yet know that whenever man talks, his speech is invariably directed to the achievement of objectives. Nevertheless, it is thus directed sufficiently often to justify our viewing speech as a likely example of a set of tools, which we shall assume to be the case.
>
> Human speech is traditionally viewed as a succession of words to which "meanings" (or "usages") are attached. We have no quarrel with this traditional view which, in fact, we adopt. Nevertheless in adopting this view of 'words with meanings' we might profitably combine it with our previous view of speech as a set of tools and stated: *words are tools that are used to convey meanings in order to achieve objectives...*
>
> Now if we concentrate our attention upon the possible internal economies of speech, we may hope to catch a glimpse of their inherent nature. Since it is usually felt that words are "combined with meanings" we may suspect that there is latent in speech both a more and a less economical way of "combining words with meanings,' both from the viewpoint of the speaker and from that of the auditor.

Information theory provides a way, unavailable to Zipf, of resolving the two viewpoints he distinguishes. The speaker's economy requires that he express messages in the most compact form possible, up to a single word that can be used in any situation for any meaning. The hearer's economy requires that the speaker be easily understandable, and thus that the amount of message reconstruction effort – including the effort of listening to the statement –be minimized. A certain minimum amount of information must be conveyed in the speaker's 'signal' so that a listener can distinguish his messages, but at the same time, too much useless information will clog things up. One can thus see that both the speaker and hearer have incentives to make the channel as efficient and easily understood as possible.

This framework can be placed on a firm mathematical footing. (Shannon 1948; Shannon 1951). Shannon analyzed all communications as a series of messages along a channel between an information source and a listener, and established mathematical bounds for the maximum amount of information, measured in bits, that can be carried along such a channel.[1] If less than this information is sent, some messages will not be distinguishable from each other. If more is sent, the "extra" is wasted resources. This is, for any source, a measure of the information content of that source and a lower bound on

the amount of time/space/bandwidth necessary to send messages from that source and be understood. And in particular, to achieve this optimum use requires a Zipf-like framework, where the most frequently sent messages have the least "information" associated with them.

This can provide a framework for mathematical analysis of the intuitive notion of language complexity. All language is a communication between the speaker and hearer; a message is "complex" if it has a large information content, and a language is "complex" if sending the message in that language requires much more bandwidth than the information content of the message.

In practical terms, there are a number of problems with the direct application of Shannon's formalism. It requires the researcher to enumerate beforehand all possible "messages" that might be sent along with their probabilities. But more serious problems lurk in the theoretical underpinnings. In the simplest formulation, Shannon's message probabilities are independent, meaning that the chance of a message being sent does not change, depending upon what other messages have been sent. In natural language, context matters, and understanding a message may depend crucially on the content of the previous messages. Finally, the assumption that the receiver may need to be able to recover the message *perfectly* might be problematic,

since the knowledge I wish to transmit may be less than the language structure demands; if all I wish to say is that one of my siblings has red hair, the sex of the relevant sibling is irrelevant, as is their comparative age, but some languages (e.g. Japanese) may force me not only to specify the sex of my sibling, but whether they are older or younger — or to engage in a long, roundabout, marked, *complex* paraphrasing. In short, the "messages" of Japanese are different than those of English, suggesting another approach may be more fruitful.

2.2 Kolomogorov and other complexity definitions

Another important measurement of complexity is that of *Kolmogorov complexity* (Li and Vitányi 1997). Kolmogorov complexity measures the informativeness of a given string (not, as in Shannon's formulation, a message source) as the length of the algorithm required to describe/generate that string. Under this formulation, a string of a thousand alternating 'a's and 'b's would be easily (and quickly) described, while a (specific) random collection of a thousand 'a's and 'b's would be very difficult to describe. For a large corpus of messages, this could be used as an operationalization of the average amount of information contained per message. In particular, notice that for

any given set of messages produced by a Shannon source, it is a very efficient use of the channel to transmit, instead of a stream of individually coded messages, an algorithmic generator for the specific stream of interest. This illustrates the close relationship between Shannon entropy and Kolmogorov complexity: Shannon's entropy is an upper bound on (and asymptotically equal to) Kolmogorov complexity. Although the mathematics required to prove this is non-trivial, the result can be seen intuitively by observing that a decompression program *and* a compressed file can be used to (re)generate the original string.

Unfortunately, Kolmogorov complexity is formally uncomputable, in a strict technical sense related to the Halting Problem. Despite this technical limitation, Kolmogorov complexity is of interest as an unattainable ideal. If, as argued above, Kolmogorov complexity represents the ultimate possible file compression, a good file compressor can be seen as an attempt to approximate this kind of complexity within a tractable formal framework. By restricting the kind of operations permitted, it may be possible to develop a useful complexity measurement framework.

One example of such a restriction is that of *linear complexity.*(Massey 1969; Schneier 1996) Linear complexity addresses the issue by assuming that

the reconstruction machinery/algorithm is of a specific form, a *linear feedback shift register* (LFSR, see figure 1) composed of an ordered set of (shift) registers and a (linear) feedback function. The register set acts as a queue, where the past few elements of the sequence line politely up in order, while the feedback function generates the next single text element and adds it to the end of the queue (dropping the element at the head, of course). The linear complexity of a given sequence is defined as the size of the smallest LFSR generating a given sequence, and can be efficiently determined by a simple computer program (Massey 1969). Such systems are widely used as cryptographic encoding systems and random number generators because of their tremendous ability to generate long, apparently complex sequences from very little information.

What are the implications of this form of complexity? In this framework, the next element of the sequence (be it a word, morpheme, phoneme, etc.) is completely determined by a deterministic function of the most recent N elements of the sequence. As will be discussed later, there is no place in this framework for the notion of context, lexicon, or even long-term memory.

Another commonly-used framework is that of *Ziv-Lempel complexity*, (Lempel and Ziv 1976; Ziv and Lempel 1977) (the complexity metric that underlies

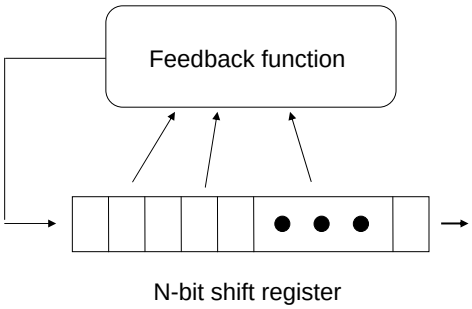

Figure 1: Sample Linear-Feedback Shift Register (LFSR)

most of the ZIP family of commercial file compressors) which by contrast involves long-term memory almost exclusively. In this framework, the system builds a collection (lexicon?) of previously-seen strings and compresses new strings by describing them in terms of previously seen items. This adaptive "lexicon" is not confined solely to traditional lexemes, but can also incorporate short-range collocations such as phrasal verbs or common combinations. It has been proven that Ziv-Lempel complexity, as instantiated in LZ77 (Ziv and Lempel 1977) or similar techniques will, given infinite computing power and memory, compress any communications stream to any desired closeness to the theoretical maximum density, without prior knowledge of the messages or probabilities of the information sources. The popularity of the ZIP program is a testament to the practical effectiveness of this model.

At least in theory, this argument (although not necessarily the psychological underpinnings) would apply to any other proposed form of file compression.

These techniques provide a method of testing and measuring the amount of information and the amount of redundancy in any string. In particular, as will be discussed in the following section, by comparing the analytic results of a complexity metric on a large set of comparable messages, one can infer

a detailed analysis not only of overall complexity, but of its linguistic and cognitive components.

3 Linguistic Experiments

3.1 Information and language

To perform this analysis, we first must consider the types of information of interest. Elsewhere (Juola 1997), it has been argued that the information relevant to a text can be broken down into four major categories :

- the complexity of the idea(s) conveyed
- the complexity of the author's style
- the complexity mandated by the language in which the author writes
- the shared information omitted between the author and her audience

The third aspect is of course what is traditionally meant by "linguistic complexity"; by holding other aspects constant and varying the language in which a particular communication occurs, this can be observed numerically as an overall language complexity metric. Neglecting for a moment the question of individual authorial style, we can compare this by measuring the

information contained in several expressions of the same ideas. A language with high "complexity" in McWhorter's sense will require more measured information in samples written in that language.

As an example, consider a simple propositional message, such as *my brother has red hair.* From a purely semantic analysis of the meanings of the words, it is apparent that a single person is being spoken of and that the person is neither myself, nor the listener. It is not necessary to transmit this explicitly, for example via third person singular verb inflection. The "information" in this inflection is therefore additional complexity demanded by the structure of the language in which it is expressed. We can further distinguish between redundancy and complexity in the nature of this information. If English (or any language) were perfectly regular, the nature of the inflection would be prefectly (and redundantly) predictable from the semantics expressed. In a more realistic setting, the inflection carries additional information necessary to describe the inflection itself (e.g., the irregularity of the verb *to have*) that is irrelevant, unnecessary, and "complex." In theory, a perfect compression program would be able to extract all the redundant information, but would still need to track the complexity.

Other examples of this sort of mandatory complexity would include gender markings, various forms of agreement, different lexical paradigms such as conjucations or declensions, and different syntactic structures such as word order, passivization, and so forth. We can continue this analysis by focusing on the information contained in specific aspects or levels of language. For example, a language that is morphologically rich would be expected to contain much of its mandatory information in the system of inflectional morphology. By systematically distorting the samples to remove inflectional morphology, one can determine the importance of this type of complexity in the overall measure.

3.2 Validating Complexity Measurements

A first requirement for any empirical test of these theories is a suitable testbed; a collection of the same messages (as far as possible) expressed in different languages. Fortunately, this material is widely available in translated literature; the Bible, in particular, is an ideal sample.(Resnik, Olsen, and Diab 1999) It is relatively large, widely available, for the most part free of copyright restrictions, and generally well-translated.

In conjunction with the definitions of the previous section, we can then

formalize common wisdom (*pace* McWhorter) regarding linguistic complexity as follows:

For any complexity metric, the measured complexity of the Bible should be the same regardless of the language or translation.

A preliminary study (Juola 1998) has provided some data in support of this hypothesis. Using Bible versions in six languages (Dutch, English, Finnish, French, Maori, and Russian), it was shown that the variation in size of the uncompressed text (4242292 bytes, +/- 376471.4, or about 8.86% variation) was substantially more than the variation in size after compression via LZ (1300637 bytes, +/- 36068.2, about 2.77%). This strongly suggests that much of the variance in document size (of the Bible) is from the character encoding system, and that the underlying message complexity is (more) uniform.

We can contrast this with the results found (Juola 2000) by a comparable study of the linear complexity of the Bible translations. In this study, it was shown that the linear complexity of the Bible varies directly with the number of characters in the relevant translation, and that there is no descriptive advantage to the linear complexity framework. (We are tempted to draw the

related conclusion that this similarly illustrates that the underlying process behind linear complexity does not describe human cognition well.)

Ideally, we would have some way of directly validating compression-based complexity measurements. One approach to this has been by setting up artificial environments such as inflectional paradigms and to vary the complexity of the paradigms. One such experiment (Juola, Bailey, and Pothos 1998) developed an artificial lexicon and systematically varied the percentage of words subject to a simple inflectional "rule," from 0 to 100%. The resulting word list(s) were subjected to an operation the researchers termed "scrampression," repeated permutation followed by compression using the LZ formulation. The results, replicated here as figure 2, show that, as expected, the "simplest" system is no inflection at all, that the system where all words were inflected was more complex (reflecting the additional complexity of the rule itself), but that the measured complexity varied smoothly and continuously, and that the most complex system (as predicted by Shannon's mathematics) was at intermediate stage where the question of whether a word was subject to the rule was itself complex.

From a mathematical and methodological perspective, these results indicate that LZ-based complexity measurements are practical to take from

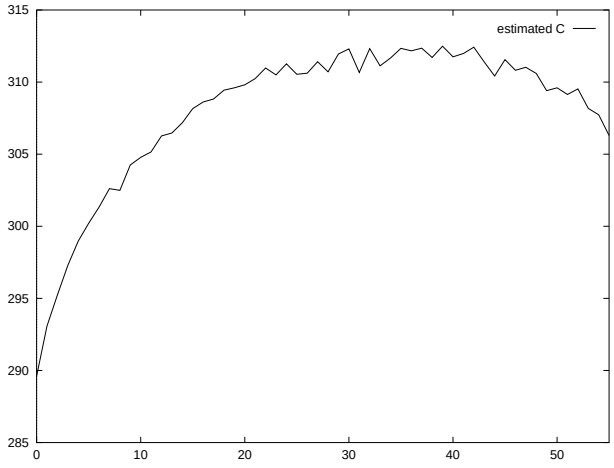

Figure 2: Variations in measured complexity as number of inflected words varies from 0 (0%) to 55 (100%)

appropriate corpora, and give meaningful measurements. Linear complexity measurements can also be taken, but do not map to our preliminary intuitions and cannot be validated against the world. From a psycholinguistic perspective, this argues strongly that LZ, with its focus on the lexicon and long-term storage and retrieval of words is a better model of the underlying regularities of language *as expressed in corpora* than linear complexity is. This is of course unsurprising; no sane person would believe a theory of language processing that did not include a lexicon. But this also suggests that careful choice of a model (and comparison of results) can illustrate other aspects of the *cognition* underlying human language, either in ways that it is the same or different from traditional, information-theory based computer programs. In particular, there are lots of other non-LZ based compression programs, many of which apparently work better on language than LZ-compression. Can we infer properties of the human mind from these programs?

3.3 Further compression analyses : Translation

The previous section discussed three positive factors in linguistic complexity in conjunction with an apparent negative aspect, "the complexity/information omitted as shared between the author and her audience." One observation

made by scholars of the translation process is that text in translation tends to be more explicit than the source material. This is easy to understand from a theoretical perspective. Information believed to be held in common between the speaker and hearer need not be expressed. For example, a writer and a reader who share a common knowledge of the city layout can be less explicit in giving directions than a writer and her out-of-town visitor.

One of the more obvious example of common knowledge is, of course, the (shared) knowledge of a language as held between two speakers. Attached to this formal knowledge, in addition, is a host of other associations, cultural cues, and general knowledge that can be broadly lumped together as "cultural context." It is reasonable to assume that two fluent speakers of the same language/dialect share a tremendous amount of associations and general knowledge that is not necessarily a formal aspect of the language; just as a physicist can omit the definition of "momentum" when talking to another physicist, a native Englishman can omit the observation that Manchester is north of London when talking to another Englishman, but not necessarily when speaking to a foreigner. Similarly, these cities' locations would probably be known to a British reader but not necessarily to the readers of the same work *in translation*. (Baker 1993) suggests that this

feature of increased specificity may be a universal aspect of the translation process and presents a brilliantly clear example, where the (English) sentence "The example of Truman was always in my mind" is translated, into Arabic, into a huge paragraph—the English translation of this paragraph containing four sentences and ninety-one words, explaining Truman's relationship to the American Presidency and the end of the Second World War, his political background, and the metaphorical relevance of this "example." This sort of implicit information need not even necessarily be "factual," but can instead be habits of thought, expectations, associations, and so forth.

If true, this conjecture and theoretical framework may allow us to measure the degree of shared knowledge between a reader and writer, even between the writer and a "general audience." For the simplest case, that of two different linguistic communities viewed as separate "general audiences," the primary difference is, of course, the social and cultural context of both. This conjecture has been tested directly (Juola 1997) using the LZ compression test defined above.

In particular, the Biblical data defined above incorporate the *original* language as one of the test set.[2] If the translation process adds information as suggested above, the Hebrew original should be significantly shorter than

Language	Size (raw)	Size (compressed)
English	859,937	231,585
Dutch	874,075	245,296
Finnish	807,179	243,550
French	824,584	235,067
Maori	878,227	221,101
Russian	690,909	226,453
Mean	822,485	235,775
Deviation	70,317	10,222
HEBREW	(506,945)	(172,956)

Table 1: Sizes and information content (in bytes) of Pentateuch

any of the translations after compression. As can be seen in table 1, this is true.

Again, we see that the variation in compressed size is much smaller than the original, and we see that Hebrew is substantially smaller than either. Of course, written Hebrew omits vowels, but we can see that this is not the primary cause of this by observing other texts.

Table 2 shows similar sizes for a variety of translations of Orwell's *1984*, originally written in English and made available in a variety of languages by the Copernicus 106 MULTEXT-East Project[3] or by the ECI/MCI corpus from the ACL. Of the nine versions of *1984*, the two smallest are the (original) English. Similarly, the smaller version of *Dark Freedom* (again distributed by ECI/MCI) is the Uzbek original.

Work	Language	Size
1984	ENGLISH (v2)	228,046
1984	ENGLISH	232,513
1984	Slovene	242,088
1984	Croat (v2)	242,946
1984	Estonian	247,355
1984	Czech	274,310
1984	Hungarian	274,730
1984	Romanian	283,061
1984	Bulgarian	369,482
Dark Freedom	UZBEK	80,209
Dark Freedom	English	90,266

Table 2: Information content of other works

We thus see that the original version of corpora tend to be smaller than their translations, in keeping with the theoretical predictions of the translation scholars.

3.4 Complexity via distortion : Preliminary experiments

With this framework in hand, we are able to address more specific questions, such as the comparative role of syntax and morphology in the complexity of a particular language. In particular, just as one could measure the complexity of a morphological rule by systematically varying the degree to which it applied, so one can distort the morphology of a language sample as a whole to see the effects of "morphological complexity."

As before, we start with some definitions. In particular, we treat the ability to predict *parts* of a word token as a measure of morphology. For example, a plural context predicts a plural noun, which in English is typically marked with the *-s* suffix. The fact that the suffix *-ing* often signals a present participle and therefore something likely to follow the string "I am" is a morphological regularity. If the noun form is suppletively irregular, then there is no *morphological* complexity involved — simply a different lexical item. A morphologically complex language, under this view, is simply one where the information conveyed by these processes contributes substantively to the information conveyed by the entire text: for example, one where the agent/patient relationships cannot be determined by examination of the word order but can be determined by inspection of (e.g.) the regularities in the endings of the individual words.

By systematically distorting these regularities, we can measure the role of morphological information within a corpus. For example, rearranging the word order would destroy syntactic relationships (one would no longer be able to identify agent and patient in English), but not morphological ones (they would still be identifiable in strongly case-marked languages). Morphological regularities can be easily hidden by simple type-substitution.

jump	walk	touch	8634	139	5543
jumped	walked	touched	15	4597	1641
jumping	walking	touching	3978	102	6

Figure 3: Example of morphological degradation process

Consider figure 3: by replacing each type in the input corpus with a randomly chosen symbol (here a decimal number), the simple regularity between the rows and columns of the left half of the table have been replaced by arbitrary relationships; where the left half can be easily described by a single row and column, the entire right half would need to be memorized individually. More accurately, this process is carried out tokenwise where each token of a given type is replaced by the same (but randomly chosen) symbol which is unique to each type. This represents a hypothetical alteration to a language where the morphological rules have been "degraded" into a situation where all versions of related words can only be treated as suppletive forms, or alternately where the morphological regularity has been drastically altered in favor of a very complex system of lexical choice.

Note, however, that this does not eliminate lexical information. If, in the original text, "I am" predicted a present participle, the new symbols that replace "I am" will predict (the set of) symbols which correspond to and replace present participles. However, because these symbols have been

randomly rewritten, there will be no easy and simple properties to determine which symbols these are. Examining figure 3 again, the third row contains the present participles. In the left hand side, all entries in this row are instantly recognizable by their "-ing" morphology; the right hand side has no similar test. In other words, there is no longer a regular (morphological) way of predicting anything about the new token. Compressing the resulting substituted file will show the effects on the "complexity," i.e. the role of morphological complexity.

Performing this substitution has shown (Juola 1998) that languages differ strongly in the way and to the degree that this substitution affects their (compressed) sizes. The results are attached as table 3. As can be seen, the resulting r/c ratios sort the languages into the order (of increasing complexity) Maori, English, Dutch, French, Russian, Finnish. It is also evident that there is significant (phonological) information which is destroyed in the morphological degradation process, as three of the six samples actually have their information content *reduced*.

Validation of these numbers, of course, might be problematic; although few would argue with any measure of morphological complexity that puts Russian and Finnish near the top (and English near the bottom), there are

Table 3: Size (in bytes) of various samples

Language	Uncompressed	Comp.(raw)	Comp. ("cooked")	R/C Ratio
Dutch	4,509,963	1,383,938	1,391,046	0.994
English	4,347,401	1,303,032	1,341,049	0.972
Finnish	4,196,930	1,370,821	1,218,222	1.12
French	4,279,259	1,348,129	1,332,518	1.01
Maori	4,607,440	1,240,406	1,385,446	0.895
Russian	3,542,756	1,285,503	1,229,459	1.04

few numbers against which to compare this ranking. Fortunately, three of the languages in this sample (English, French, and Russian) are also part of the sample addressed in (Nichols 1992). Nichols finding is that English is *less* complex than either Russian or French, which are themselves equivalent. The ranking here agrees with Nichols' ranking, placing English below the other two and French and Russian adjacent. This agreement, together with the intuitive plausibility, provides a weak validation.

Further examination of the corpora indicates some other interesting findings. These findings, presented as table 4, can be summarized as the observation that (within the studied samples) the ordering produced by the complexity-theoretic analysis is identical with the ordering produced by the number of word *types* in the samples, and identical-reversed with the ordering produced by the number of word *tokens*. (This identity is highly significant; Spearman's rank test yields $p < 0.0025$, while Kendall's T test yields $p <$

Table 4: R/C ratios with linguistic form counts

Language	R/C	Types in sample	Tokens in sample
Maori	0.895	19,301	1,009,865
English	0.972	31,244	824,364
Dutch	0.994	42,347	805,102
French	1.01	48,609	758,251
Russian	1.04	76,707	600,068
Finnish	1.12	86,566	577,413

Table 5: Type/Token ratio across cased/uncased languages

Concept/Case	English	Latin
Nominative	(the) night	nox
Genitive	of (the) night	noctis
Dative	to (the) night	nocti
Accusative	into (the) night	noctem
Ablative	in (the) night	nocte
Number of types	5(6)	5
Number of tokens	9(14)	5

0.001.) In other words, languages which are morphologically complex tend to have a wide variety of distinct linguistic forms, while languages which are morphologically simple have more word tokens, repeating a smaller number of types, as in table 5.

This finding is intuitively plausible; the linguistic constructs which in a language like Russian or Latin are expressed in terms of morphological variation are expressed in other languages through function words—which almost by definition are few types but many tokens. However, this finding is not *necessary*; it is theoretically possible for a language to exist that makes

use of an extraordinarily wide variety of function word types (thus inflating the number of types) or that inflates the number of tokens (for example by indicating plurality with repeated tokens). This finding then is further evidence for the approximate equality of overall linguistic complexity, at least within this sample.

3.5 Complexity via distortion : Subsequent experiments

A larger-scale version of these results (Juola 2005) confirms these findings. An analysis of twenty-four Bibles, including nine English translations and sixteen non-English versions, shows, again, languages are about equally complex, but that they express their complexity differently at different levels. (Table 6 shows the languages represented in this study.)

Methodologically, this study was slightly different, both the scope of the levels studied, but also in how distortion was performed. All samples were divided identically into verses (e.g. Genesis 1:2 represented the same "message" as best the translators could manage). Samples were distorted morphologically by the deletion of 10% of the letters in each verse at random. Samples were distorted syntactically by the deletion of 10% of the words (maximal non-blank sequences), while pragmatic distortion was obtained by deleting

English Versions	Non-English Versions
American Standard Version (asv)	Bahasa Indonesian (bis)
Authorized (King James) Version (av)	Portuguese (brp)
Bible in Basic English (bbe)	Haitian Creole (crl)
Darby Translation (dby)	French (dby)
Complete Jewish Bible (jps)	Finnish (fin)
Revised Standard (rsv)	Hungarian (karoli)
Webster's Revised (rweb)	Dutch (lei)
Young's Literal Translation (ylt)	German (lut)
	Modern Greek (mgreek)
	Modern Greek [unaccented] (mgreeku)
	French (neg)
	Russian (rst)
	German (sch)
	German (uelb)
	Ukranian (ukraine)

Table 6: Bible samples and languages used

10% of the verses themselves at random. We justify this as an exploration of "pragmatic" complexity by noting that sentences themselves must be interpreted through context. For example, the use of pronouns hinges on the availability of antecedents; languages where subjects can be dropped altogether make even heavier use of context. Just as elimination of words can distort the structure of individual sentences and make them difficult to follow, so eliminating previous sentences or verses, while having no effect on the syntactic acceptability of the individual sentences, will make the *pragmatics* of the discourse difficult to follow.

Although all three procedures will provably delete the same number of

expected characters from the source files, the effects of compression on these files was expected to differ as argued above. Compression was performed using the UNIX *gzip* utility. [4]

Table 7 shows the compressed and uncompressed sizes (in bytes) of the various Biblical translations; Table 8 presents the (compressed) sizes both of the original and of the various distorted versions. Table 9 normalizes all results by presenting them as multiples of the original size (e.g. the morphologically distorted version of 'asv' is 1.17 times its original size when both are compressed). This table is also sorted in order of decreasing value, and English versions are labeled in ALL CAPS for ease of identification. Examining these tables together supports the hypotheses presented earlier. First, all languages appear to be more uniform in size in their compressed than in their uncompressed sizes. Second, languages appear to differ in a reliable way in the degree to which their complexity is expressed in morphology or syntax; languages with high measured morphological complexity (as measured via morphological distortion) have low measured syntactic complexity (as measured via syntactic distortion). A new finding of this experiment is that all languages appear to be about equal in their *pragmatic* complexity; the variation in column 3 of table 9 is almost nonexistent. This suggests that

Version	Uncompressed	Compressed
asv	4280478	1269845
av	4277786	1272510
bbe	4309197	1234147
bis	4588199	1346047
brp	3963452	1273549
crl	4351280	1312252
dby	4227529	1265405
drb	4336255	1337768
fin	4169287	1370042
jps	4309185	1288416
karoli	3957833	1404639
lei	4134479	1356996
lsg	4252648	1347637
lut	4180138	1341866
mgreek	4348255	1468263
mgreeku	4321200	1341237
neg	4188814	1323919
rst	3506401	1269684
rsv	4061749	1253476
rweb	4247431	1262061
sch	4317881	1417428
uelb	4407756	1383337
ukraine	3564937	1315103
ylt	4265621	1265242

Table 7: Uncompressed and compressed sizes (bytes) of various Biblical translations

conversation-level constructions are handled similarly in all languages studied, perhaps reflecting underlying similarities in overall cognitive processing.

One minor point needs to be addressed. It is intuitively plausible that deleting words or phrases from a corpus should result in the overall lowering of information. It is less plausible that deleting letters should result in the overall increase in information — but that is a clear result of table 9. The explanation is simple, but unhelpful. Essentially, by deleting letters randomly, the computer is creating a number of variant spelling of the same word, or alternatively a number of freely varying synonyms for each lexical item. Unsurprisingly, the compression program needs to determine (and store) which of the variant spellings is used in each instance. Strongly inflecting languages such as Finnish and Hungarian, ironically, suffer the least penalty because they have more word forms, and therefore fewer new variants are introduced in each one. Thus, the results show both that Finnish and Hungarian have high morphological complexity but low syntactic complexity, a fact that is both intuitivel and supported by prior experiments.

It should finally be noted that in this sample, there is a single example of a creole language present. Sample 'crl' is Haitian Creole. Under McWhorter's hypothesis, this should be the smallest in compressed size. It is, in fact, about

Ver..	Normal	Morph.	Prag.	Syn.
asv	1269845.00	1487311.21	1162625.76	1235685.14
av	1272510.00	1483580.35	1164269.82	1236878.18
bbe	1234147.00	1461058.00	1130314.66	1207880.69
bis	1346047.00	1601727.60	1229751.14	1282227.44
brp	1273549.00	1452517.41	1164547.00	1226345.93
crl	1312252.00	1518413.44	1200910.23	1285958.47
dby	1265405.00	1474721.85	1158317.63	1229267.18
drb	1337768.00	1557772.69	1224481.67	1295321.34
fin	1370042.00	1548445.81	1251576.37	1294915.38
jps	1288416.00	1504061.39	1179304.44	1253465.97
karoli	1404639.00	1575233.97	1283223.98	1330816.55
lei	1356996.00	1522644.85	1239991.16	1297103.54
lsg	1347637.00	1560146.19	1233249.90	1300432.48
lut	1341866.00	1529869.88	1226982.62	1285579.01
mgreek	1468263.00	1701221.45	1343161.70	1408603.16
mgreeku	1341237.00	1550600.30	1226936.07	1292009.44
neg	1323919.00	1533468.52	1211325.57	1277609.72
rst	1269684.00	1410919.44	1161101.89	1204901.46
rsv	1253476.00	1445795.94	1147694.76	1215960.88
rweb	1262061.00	1471718.69	1154952.77	1227624.14
sch	1417428.00	1604828.85	1295267.52	1353695.77
uelb	1383337.00	1598355.09	1265317.06	1329331.17
ukraine	1315103.00	1451550.59	1201829.66	1247197.05
ylt	1265242.00	1482038.11	1158189.17	1230766.62

Table 8: Biblical sizes (bytes) after various transformations

Ver..	Morph.	Prag.	Syn.
ukraine	1.10375	0.913867	0.948365
rst	1.11124	0.914481	0.948977
karoli	1.12145	0.913561	0.947444
lei	1.12207	0.913777	0.955864
fin	1.13022	0.913531	0.945165
sch	1.13221	0.913815	0.955037
lut	1.14011	0.914385	0.958053
brp	1.14053	0.914411	0.962936
RSV	1.15343	0.91561	0.970071
uelb	1.15543	0.914685	0.96096
mgreeku	1.1561	0.914779	0.963297
crl	1.15711	0.915152	0.979963
lsg	1.15769	0.91512	0.964972
neg	1.15828	0.914954	0.965021
mgreek	1.15866	0.914796	0.959367
drb	1.16446	0.915317	0.968271
DBY	1.16541	0.915373	0.971442
AV	1.16587	0.91494	0.971999
RWEB	1.16612	0.915132	0.972714
JPS	1.16737	0.915313	0.972874
ASV	1.17125	0.915565	0.973099
YLT	1.17135	0.915389	0.972752
BBE	1.18386	0.915867	0.978717
bis	1.18995	0.913602	0.952587

Table 9: Normalized sizes (bytes) after various transformations

average. Readers who wish to regard this as evidence against McWhorter's hypothesis are sharply reminded of the dangers of drawing inferences from a single data point. As discussed in the next section, a much larger experiment on this scale should be run.

4 Discussion

The results presented above can be viewed as an extended proof of concept for the idea that "language complexity" is a meaningful concept subject to empirical testing. In particular, the terminology and concepts of information theory can provide a structure for developing such studies. Taking a functionalist view of language — language serves to pass a message between the speaker and hearer — one can compare the mathematical properties of several different expressions of the same message.

In particular, by focusing on different translations of "the same" text, we can establish whether there are underlying differences in these properties that characterize the language in which the texts are written. By systematically distorting a single base text, we can establish whether this distortion introduces systematic changes in the properties.

The key questions are thus: what are the (mathematical) properties that

can be measured, and how do they relate to the (linguistic) properties of general interest? In particular, how do our intuitive understandings of "complex" relate both to language and to mathematics?

The variety of definitions of complexity, both linguistic and mathematical, suggest that this relationship is not well-understood within either community. One possible approach to resolving this is to examine more closely the notion of process and the relationship of the processes captured in both psycholinguistics and the mathematical formalisms. In particular, most complexity formalisms start from a set of mathematical or algorithmic "primitives," and describe the complexity of an object in terms of the number of primitives required for it to function. For example, the fundamental primitives underlying Kolmogorov complexity are exactly that underly Turing machines — reading and writing unanalyzed symbols in an unbounded memory space. The fundamental primitives of LZ complexity include storage and retrieval from a learned "lexicon" of expressions. The "complexity" is defined as the number of primitives in a description of the process, ignoring completely the time element (how many primitives need actually be *performed*). By contrast, linear complexity assumes a very strong bound on the amount of memory available, and measures the complexity by the size of this very bound. Since

human memory is known to have a large capacity, it makes intuitive sense that linear complexity should not well measure aspects of human cognition – while the plausibility of the lexicon strengthens the plausibility of LZ-based complexity measurements..

The framework developed in this paper thus provides researchers with a new way of investigating theories of the psycholinguistic processing of language. Specifically, by identifying the theoretical "primitives" implicit in the psychological theory, and developing a text compression method based upon those primitives, the results of compressing should be at least as plausible and at least as successful as the results obtained by other compression methods such as LZ-compression.

Similarly, a good theory of language processing should be able to capture the underlying regularities of human language in a sufficiently practical way that will be useful for the quite practical problem of text compression. With the ever increasing amount of text available (Nerbonne 2004), the ability to store ideas in a way that matches the efficiency of the human mind is almost certainly a desirable goal.

Finally, examination of algorithmic representation in terms of linguistic primitives may provide medical and psychological advantages as well. If our

theory represents human performance under ideal conditions, studying how performance changes when theory is tweaked slightly may illustrate or explain aspects of human performance under less than ideal conditions, including both situational and cognitive degradation. This may improve science's ability both to diagnose and to treat linguistic/cognitive disorders.

5 Future Work and Conclusions

Briefly to recap the findings of the previous sections: the question of "linguistic complexity" can be formalized objectively and with relative precision using information theory. Depending upon the exact definition of "complexity" used, different measurements will be obtained. In particular, by using multiple translations of the same basic text (for example, the Bible or the novel *1984*), one can see whether "all languages are about equally complex" as received wisdom requires, whether a particular language or language group is systematically less complex as McWhorter suggests, and whether there any interesting variation or pattern in complexity.

Performing this experiment shows that languages do appear to be about equally complex when compared under the Ziv-Lempel definition of complexity. Less psycholinguistically plausible definitions, such as linear complexity,

show correspondingly less plausible results. In particular, the variation in text lengths of uncompressed versions of the Bible was significantly greater than the variations in compressed lengths.

Further studies showed that varying particular aspects of language complexity — the regularity or irregularity of morphology — could similarly be detected by compression-based information theory measurements. Well-known aspects of language-in-translation, specifically the observation that translated texts need to be more explicit so that the reader can understand them against a different cultural background, can be observed by noting that translated texts are more complex, in this framework, than their originals. Finally, by systematically distorting particular types of expression, one can measure the role that type of expression plays in the overall complexity of language. Using this framework, studies have shown that languages differ in their morphological and/or syntactic complexity, and in particular, that languages with high morphological complexity have low syntactic complexity and vice versa. This is compatible with the tool-and-effort framework of Zipf and of Shannon; a language capable of sending information via morphology need not encode the same information syntactically A further finding — that languages do not appear to differ in their pragmatic complexity — may

reflect the underlying cognitive universals of the human mind/brain system.

These findings only scratch the surface of what could be explored under this framework. In particular, McWhorter's theory that creoles are less complex than older languages could be tested by collecting a corpus of texts written in a variety of languages including both creoles and non-creoles. More carefully, one would want a text sample, written in one language and independently translated into a second (creole) and third (non-creole) language. Assuming that the same concepts are expressed in each translation, any significant size differences could only be attributed to the structural properties of the languges. If the creole translations were systematically smaller than the non-creole ones, this would be strong evidence supporting McWhorter's theoretical observations.

Another area of immediate research could be in the study of other formalizations of complexity and their usefulness and plausibility as linguistic measurements. The three concepts presented above are only a few of the dozens of definitions proposed and used by information theorists as compression technology. A systematic study of how different measurements vary would help researchers begin to understand which methods give intuitively plausible results and which don't.

Beyond this, the primary difference between the intuitively plausible results of the Ziv-Lempel metric and the implausible ones of linear complexity have been argued to be a result of the difference in handling of long-term memory and the lexicon. Can we find other methods that illustrate or demonstrate other important aspects of cognition?

This is the ultimate promise of complexity studies. A recent problem in linguistics has been dealing with the results of corpus-oriented studies; the problem of inferring linguistic *processes* from simply collections of *products*. The process-oriented analysis implicit in information theory may provide a useful tool to help in this inference and bridge the gap between the paper and the mind.

Notes

[1] More formally, Shannon demonstrated that, given a source S, capable of sending any of N messages $i = 1, \ldots, N$, the average length of a message (in bits), is given by

$$H(S) = \sum_{i=1}^{N} p_i \cdot l_i \tag{1}$$

where p_i is the probability that message i is sent, and l_i is the length of message i. The optimum lengths are when l_i is equal to the negative logarithm of p_i. Thus,

$$H(S) = \sum_{i=1}^{N} -p_i \log_2(p_i) \tag{2}$$

[2]The Bible as a whole is actually somewhat problematic in this regard as there is no single original language. For this reason, we momentarily restrict our attention to the first five books, which are all originally written in Hebrew.

[3] This project is described at `http://nl.ijs.si/ME/Corpus/mte-D21M/mte-D21M.html`

[4]Another analysis, using a slightly different compression scheme (*bzip2*, using the Burrows-Wheeler transform), obtained similar results but will not be discussed further.

References

Baker, M. (1993). Corpus linguistics and translation studies : Implications and applications. In M. Baker, G. Francis, and E. Tognini-Bonnelli (Eds.), *Text and Technology : In Honour of John Sinclair*, pp. 233–250. Amsterdam: John Benjamins.

Juola, P. (1997). A numerical analysis of cultural context in translation. In *Proceedings of the Second European Conference on Cognitive Science*, Manchester, UK, pp. 207–210.

Juola, P. (1998). Measuring linguistic complexity : The morphological tier. *Journal of Quantitative Linguistics 5*(3), 206–13.

Juola, P. (2000, August). A linear model of complexity (and its flaws). In *Fourth International Conference on Quantitative Linguistics (QUALICO-2000)*, Prague, Czech Republic.

Juola, P. (2005, September 24–26). Compression-based analysis of language complexity. In *Presented at Approaches to Complexity in Language*, Helsinki, Finland.

Juola, P., T. M. Bailey, and E. M. Pothos (1998, August). Theory-neutral system regularity measurements. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society (CogSci-98)*, Madison, WI.

Lempel, A. and J. Ziv (1976, January). On the complexity of finite sequences. *IEEE Transactions on Information Theory IT-22*(1), 75–81.

Li, M. and P. Vitányi (1997). *An Introduction to Kolmogorov Complexity and Its Applications* (2nd ed.). Graduate Texts in Computer Science. New York: Springer.

Massey, J. L. (1969, January). Shift-register synthesis and BCH decoding. *IEEE Transactions in Information Theory IT-15*(1), 122–7.

McWhorter, J. H. (2001). The world's simplest grammars are creole gram-

mars. *Linguistic Typology 6*, 125–166.

McWhorter, J. H. (2005). *Defining Creole*. Oxford: Oxford University Press.

Nerbonne, J. (2004, June). The data deluge. In *Proc. 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*, Göteborg, Sweden. To appear in *Literary and Linguistic Computing*.

Nichols, J. (1986). Head-marking and dependent-marking grammar. *Language 62*, 56–117.

Nichols, J. (1992). *Linguistic Diversity in Space and Time*. Chicago, IL: University of Chicago Press.

Resnik, P., M. B. Olsen, and M. Diab (1999). The Bible as a parallel corpus: Annotating the 'Book of 2000 tongues'. *Computers and the Humanities 33*((1-2)), 129–153.

Schneier, B. (1996). *Applied Cryptography, Second Edition : Protocols, Algorithms and Source Code in C*. New York: John Wiley and Sons, Inc.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal 27*(4), 379–423.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal 30*(1), 50–64.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. New York: Hafner Publishing Company. Reprinted 1965.

Ziv, J. and A. Lempel (1977, May). A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory IT-23*(3), 373–343.