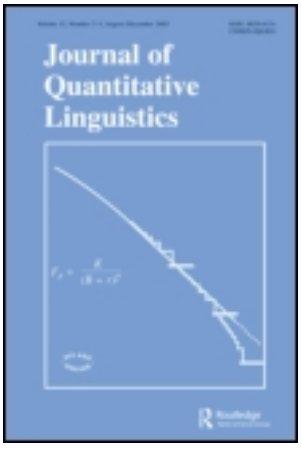

Journal of Quantitative Linguistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/njql20

Measuring linguistic complexity: The morphological tier

Patrick Juola [a]

[a] Department of Experimental Psychology, Oxford University, Oxford, OXl 3UD, UK Phone: +44 1865 271 404 E-mail:

Available online: 21 Jul 2008

To cite this article: Patrick Juola (1998): Measuring linguistic complexity: The morphological tier, Journal of Quantitative Linguistics, 5:3, 206-213

To link to this article: http://dx.doi.org/10.1080/09296179808590128

Journal of Quantitative Linguistics
1998, Vol. 5, No. 3, pp. 206–213

0929-6174/98/0503-0206$12.00

Measuring Linguistic Complexity: The Morphological Tier *

Patrick Juola
Department of Experimental Psychology, Oxford University, UK

ABSTRACT

How to measure morphological complexity is itself an issue of some complexity. (Nichols, 1992)

This article develops an information-theoretic and functionally motivated method of measuring "linguistic complexity" from corpora that can in theory be applied to any definable substructure. This method is further extended into a way of objectively and numerically measuring the morphological complexity of a given language sample, avoiding the typical difficulties of focusing on only a few unrepresentative types of constructions. By selectively altering the morphological information present in a sample, the complexity can be measured as the change in overall informativeness of a text. This claim has been tested in a small-scale cross-linguistic experiment; the results agree well with both intuitions and existing measurements.

INTRODUCTION AND BACKGROUND

A standard question in freshman linguistics is "Which language is the most complex?", with related questions about the various aspects of complexity, typically the ones the student is having trouble learning. Aside from its armchair interest, it can also be an important typological question in the description of language categories, just as the universals proposed by Greenberg (1966) have proven to be. For example, the categorisation of languages by the number of persons represented in the pronoun and verb conjugation systems is a simple categorisation about the complexity of pronominal expression. Answering this question might also shed light on the human brain's processing of linguistic information and the limitations on what sort of linguistic structures and expressions are learnable – is there an underlying reason why languages with trial or paucal pronouns are (relatively) rare, and to what other factors does this observation relate? Unfortunately, there is no accepted method for measuring and comparing aspects of complexity, and linguists are reduced to answers based more on politics and collections of ad hoc observations than on empirical evidence. Received wisdom states that, because languages are equally expressive (itself a proposition deserving of further investigation), they are equally complex, but merely express their complexity in different areas. The perceived richness of the case systems of Latin or Greek make them appear complex until one looks closely at the rigidity and "complexity" of English syntax.

On the other hand, this answer is somewhat unsatisfying from a scientific point of view. Not only would it be a coincidence of the highest order if the syntactic complexity of English *exactly* matched the morphological complexity of Greek, but it's not clear how this finding could be validated or disproven. To compare, one must first measure, and the process of measurement is not intuitively obvious.

Morphological complexity, in particular, is an obvious testbed for any theories about the

*This work was supported by ESRC grant 70. The author would also like to acknowledge the valuable help of and discussions with Todd Bailey, Lise Menn, John Reynolds, and the attendees of QUALICO-97.

Address correspondence to: Patrick Juola, Department of Experimental Psychology, Oxford University, Oxford, OX1 3UD, UK. Tel: +44 1865 271 404. E-mail: patrick.juola@psy.ox.ac.uk

possibility of getting these measurements; it is intuitively apparent that some languages (for example, Finnish) are "morphologically complex" while others are more simple. On the other hand, claims about (for example) semantic differences are less intuitive and less widely accepted.

The historical difficulty with any complexity measurement is one of scope. For instance, two classic analyses deal with basic colour terms (Berlin & Kay, 1969) and deixic complexity (Perkins, 1992). Berlin and Kay (1969) discovered an elegant hierarchy of the number of basic colour terms in various languages. It is relatively simple to determine where languages fall on this hierarchy and thus to compare the complexity of their "color systems". Unfortunately, these measurements are difficult to extend; for example, it is difficult to predict other aspects of English from the observation of how many basic colour terms it has in comparison with another language. Similarly, it is difficult to make larger predictions from observations of a language's deixic complexity, especially when these are categorised instead of measured numerically. On the other hand, it may be possible to predict, for instance, that languages with "looser" syntax would have a richer and more complex set of morphology, or that languages with a low degree of morphological complexity will have a large number of function words.

The notion of syntactic complexity as an explicitly computational concept was addressed in detail by Berwick and Weinberg (1984). Using an algorithmic parser, they derive and describe several properties of transformational grammar in terms of what are known to be efficient properties of Turing Machine calculations. The idea of efficient comparison of complex systems is implicit in the "Principles and Parameters" approach to language acquisition (Dorr, 1993; Lightfoot, 1991; Morgan, 1986) where (syntactic) language variation is explained as learned variation via a fixed set of parameters; by enumerating the parameters, one can describe a language as closer to or further from the basic grammar, and thus more "complex", although this again lacks sufficient scope to explain how a language's complexity can be balanced between morphology and syntax.

One of the most comprehensive works on morphological complexity is that of Nichols (1992). In her work, she develops a measure of complexity based on the number of points at which a typical sentence is capable of receiving inflection. This is one of a very few, and perhaps the only, attempt to produce a comparative numerical index of complexity. She further applies this analysis to a sample of nearly 200 languages, some findings of which are replicated below. It is significant that she does not attempt to justify this index in mathematical terms, since there is no well-accepted standard against which to compare this.

A further difficulty of Nichols' approach is that she measures against an idealized language involving "a typical sentence". First, this ignores the possibility of a different amount of morphological variation at a given point – for example, two languages may inflect (morphologically) for "gender", but the first language has only two genders while the second has six or more. It would be reasonable in this case to claim that the second is more complex. A more subtle difficulty lies with finding examples of "typical". From the standpoint of a language learner or user, a language with six "genders", but of which four are almost never used, is little-if-any more complex than the two-gender example above. The case for "typical" is a difficult one, especially as few examples of language can be regarded as "typical". It is into this mess that one tries to develop a theory to support numerical validation of this sort of measurement.

A PROPOSED APPROACH

The approach developed here uses a teleological framework describing language as a goal-directed, functional activity. Speakers use language to a purpose, perhaps to persuade the listener to take some action or to achieve some understanding, and have some incentive to take the least amount of effort consonant with this. (This idea is not especially new and can be traced back at least as far as Zipf, 1949.) This goal-directed process can be most clearly seen in written communication; from some background information

shared by both writer and reader, the writer produces some form of linguistic expression that, by passing information to the reader, will achieve the writer's goal. Obviously, as the communication is one-way and unrepeated, the only way the writer has to persuade the reader is through the writing – and just as obviously, the writer would rather write a hundred words than a hundred thousand, all other factors being equal.

This effort can be approximated in a measurement of the amount of information expressed in the writing of interest, as expressed below. Deferring that discussion for a few paragraphs, it is further apparent that there are several factors that contribute to this effort. The more complex and difficult the desired goal, the more effort required to describe it (thus one would expect the instructions to land a plane to be more complex than those to program a VCR). Similarly, the writer's style, skill, and estimation of his audience will influence the degree of expressivity (variance in this with regard to expressivity in translation is explored in Juola (1997)). Finally, if the equal complexity hypothesis is *not* correct (even if it's a good approximation), the language used will influence the information contained in a given communication.

However, to the degree that these factors can be controlled for, it is reasonable to ask how much of this information is expressed by morphological (as opposed to syntactic, semantic, etc.) processes. A simple example may suffice to illustrate these tiers. The English sentence "John gave a ring to the sister of Marge" describes a single giving-event, primarily in the use of the lexeme *give*. Similarly, the participants in the event, John, Marge, someone's sister, and a ring, are described by lexical choice. On the other hand, the roles of the participants are described by the (syntactic) order in which they appear in the sentence. On the other hand, a similar sentence "John gave a ring to Marge's sister" describes a similar event, with similar participants, but the role played by Marge (of an owner or possessor) is obvious from the morphological structure of the token "Marge's". In a strongly case-marked language such as Latin, the entire role system could be reproduced from the morphological information alone, indicating that there is a large(r) amount of morphological information present in Latin. This "morphological tier" is the subject of the current investigation; to what extent is the morphological information necessary to an understanding or a reconstruction of the original corpus? By selectively altering the expression of *morphological* information, one can measure the amount of morphological complexity contributes to a corpus by measuring the change in perceived informativeness. The inspiration for this measurement derives from information theory.

The mathematics of information theory (for more details, consult Abramson, 1963; Shannon, 1948) can be summarised in the basic idea that "information" equates to the unexpectedness of a piece of information and to the degree to which something cannot be simply predicted from other aspects of the sample. The most common model for measuring "information" is based on the assumption of a source of "messages" (usually treated as stochastic) transmitting over a channel to a listener. If the probability distribution for a given source is known exactly, then it's fairly easy to devise an encoding that achieves maximum transmission efficiency, described by the equation

$$H = \Sigma\, p_i \cdot \log p_i, \qquad (1)$$

where p_i is the probability of the ith event (or the ith message) occurring. This efficiency is termed "entropy" by Shannon. In the more realistic case where the probabilities are not known exactly but are only inferred from samples, adaptive methods such as described in Ziv and Lempel (1977, 1978) can approximate this maximum and therefore indirectly provide a reasonably accurate estimate of the entropy or informativeness of such a channel – the more predictible and less informative the channel, the greater a difference between the normal and maximum channel efficiency. (This process of reducing a communication channel to maximum efficiency is more commonly known as compression, and in fact most modern file compression programs use some variant of adaptive entropy estimation.)

A related measurement of complexity is that of Kolmogorov or Chaitin complexity (Li & Vitányi, 1997). Kolmogorov complexity measures the informativeness of a given string (not a source) as the size of the algorithm required to describe/generate that string. This can be used as an operationalization above of the "amount of effort" required to generate a corpus defined above. 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. A more complex string (in the Kolmogorov complexity sense) will be less compressible, and thus require a larger program *and* compressed text system to reconstruct.

METHODS

We thus have all the pieces necessary for an information-theoretic investigation of the complexity of the "morphological tier." Using standard compression techniques (Ziv & Lempel, 1977), we can reduce language samples (in the forms of ASCII-encoded corpora) to an approximation of their Kolmogorov complexity and measure the "informativeness" of the original corpora. If the hypothesis that "all languages are equally complex" is correct, the informativeness of various translations of a single work should be approximately the same and their compressed texts should be approximately the same size. It is thus possible to meaningfully compare linguistic complexity at the level of translations of individual texts. This comparison can be extended by developing and measuring the complexity of related texts with various degrees of change and/or degradation of the information contained therein.

This degradation process can be specifically operationalized to produce a definition of "the morphological tier". Morphology, for the purpose of this work, is the process by which a hearer can predict parts of a word in the input corpus, or alternately, that level of meaning which is expressed in regular (and hence predictable) relationships between groups of words. Words, for this purpose, are operationalized to maximal nonblank sequences of characters. (Obviously, this naive definition of words restricts the applicability of this technique to languages that are written with spaces separating words, and allows possible confusion between homographs.) Thus, 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. 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 words in the sentences.

These regularities can be easily hidden by simple type-substitution. Consider Figure 1: 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 memorised individually. More accurately, this process is carried out typewise 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 version of related words can only be treated as suppletive forms, or alternately where the morphological regularity has been drastically altered in favour of a very complex system of lexical choice.

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

Fig. 1. Example of morphological degradation process.

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 1 again, the third row contains the present participles. In the left hand side, all entries in this row are instantly recognisable 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.

This rewriting process will have two main effects. First, information (such as pronunciation) at the phonological tier is irrevocably destroyed; this results in a net loss of information and corresponding overall sizes of compressed files. Second, relationships between and among words at the syntactic tier (and above) are unchanged; the primary effect is to make the prediction of particular word-forms *on the basis of other word-forms in the sentence* more difficult; i.e., to inflate the information content at the morphological tier. In a "morphologically complex" language, the primary difficulty is to predict the exact word tokens (rather than, for example, token ordering), and so a language sample which is already highly informative at that tier (i.e., has words with meaningful internal structure) will not find its compressibility substantially reduced (and may find its compressibility increases, i.e., the compressed degraded text is much smaller), as the prediction of word tokens is already a difficult task. A language sample with an uninformative morphological tier, on the other hand, will have significantly more random noise and hence be significantly less compressive. By comparing ratios (for different languages) of the information contained in the raw samples with the information contained in the morphologically degraded samples one can achieve an approximate numerical measurement of morphological complexity.

To briefly recap, the above predicts that by inflating the information content of the morphological tier, languages with a regular, informative, morphology will have their information content greatly increased (i.e., higher entropy, reduced compressibility, less order and predictability) relative to the information contained in the raw, unaltered sample. Expressed numerically, one can take the ratio of information in the raw sample with the information in the "cooked" (morphologically degraded) sample, with a high ratio predicted to correspond to languages with high morphological complexity. This prediction can be seen in Fig 2. This figure shows the ratio differences between two hypothetical languages, *M* and *S*, where most of the complexity of *M* is at the morphological tier, while most of the complexity of *S* is at the syntactic tier. By artificially raising the morpholog-

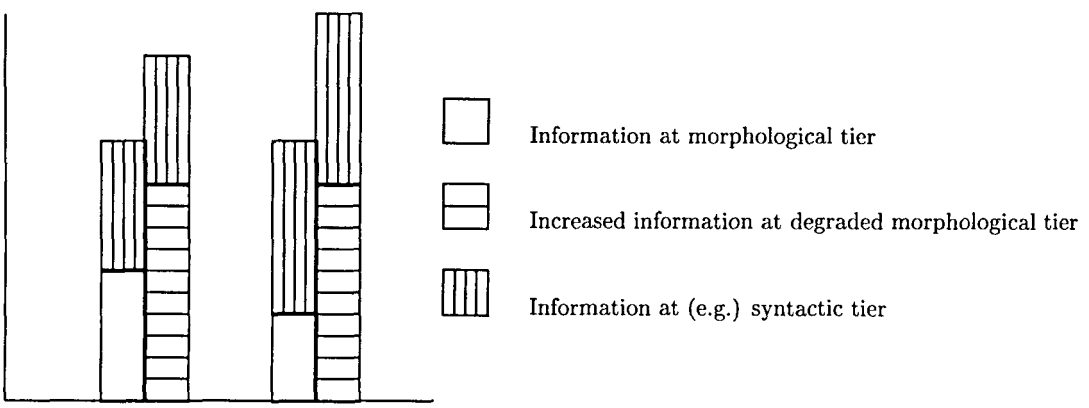

Fig. 2. Hypothetical example of degradation ratios.

ical complexity of both languages (in the second column of each pair), one can see that the ratio between the original language and the altered language (here expressed as the ratio of length of the two bars) is greater in the case of *M* than of *S*.

This conjecture can easily be tested by the method described above. One takes large samples of text in a variety of languages and uses a state-of-the-art compression algorithm to determine the amount of information contained therein; simple text transformations yield an appropriately degraded sample for comparison and elementary arithmetic will then give the appropriate ratios to be used for comparison. There is not even any particular theoretical reason that the same base text be used for the various comparisons, as the appropriate ratio may be (in theory) text-independent, but using the same underlying text will also produce results that can be more directly compared.

RESULTS

The experiment described above has been performed using Biblical texts. Pierce (1993) provides machine-readable copies of the Bible in a variety of languages and translations, including Dutch, English, Finnish, French, Maori, and Russian. The sample taken from each language was the entire text of the Bible (Old and New Testaments, but excluding Apocrypha), approximately 825,000 words in English or about 4.3 megabytes of raw text. The compressed sizes varied from 1.24 megabytes (in Maori) to 1.38 megabytes (in Dutch), a substantial reduction in the overall variance and implicit evidence for the equal complexity hypothesis.

Morphological degradation was done by simply replacing every maximal nonblank sequence with a randomly chosen number between one and the number of (distinct) sequences; e.g., one and approximately thirty-one thousand in the case of English. These degraded texts, as well as the original (raw) texts, were compressed using *GZIP* version 1.2.4, and the size of the resulting files was used as a measure of their complexity.

The results are shown in Table 1. 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.

DISCUSSION

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 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; clearly much additional work (in additional languages and/or samples) can be done to test this claim.

Further examination of the corpora indicates some other interesting findings. These findings, presented in Table 2, can be summarised as the observation that (within the studied samples) the ordering produced by the complexity-theoretic

Table 1. Size 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

Note. Size is in bytes.

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 $p < 0.001$.) In other words, languages which are morphologically complex tend to have a wide variety of distinct linguistic forms that appear in large samples, while languages which are morphologically simple have more words.

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. Instrumental relationships, for instance, are usually expressed in English using prepositions ("opening a door *with* a key"), but in Russian with a morpheme attached to the instrument. Benefactives ("donating a painting *to* a museum") can show a similar pattern. Similarly, a language with a rich morphological inflection system is capable of producing a wide variety of types, and thus it's not surprising that it would use more of them. However, this finding is not *necessary*, in the sense that 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.

Table 2. *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

CONCLUSIONS AND FUTURE WORK

The findings above are largely unsurprising; few would quibble with the statement that some languages are more morphologically complex than others, or that Finnish and Russian are complex compared to English. From an empirical standpoint, however, these results are interesting for several reasons. First, they provide evidence in support of the usefulness of a general information-theoretic approach to linguistic complexity which may be extensible to other areas. Second, they demonstrate an empirical and objective approach to the direct measurement of something (morphological complexity) previously subjective. The existence of a numerical approach can only be helpful, even if all it does is help refine the original question of "what is linguistic complexity?" and determine to what extent (for example) number of linguistic types is a valid measure of such complexity.

This method of selective alteration (or deletion) of tiers can presumably be extended to other areas; for example, by selectively altering the syntactic tier, one could produce similar measures for the syntactic complexity of a given language. This would have a clear advantage in syntactic complexity measurement over generalisations of Nichols' technique (counting the points in a sentence where additional phrases could be added?) or type/token ratios (as most sentence types are unique, the ratio would almost always be one). Similarly, it may be possible to investigate the complexity of sublanguages – one would intuitively expect that (English) scientific jargon, with its wealth of Latinate compositional morphology, would be "more complex" than mainstream fiction. Much further work is clearly required, both to refine the tools used and to more carefully measure their accuracy, as well as to determine the areas and extent of their usefulness. This work only touches the surface of what may prove to be a very wide-ranging and fruitful area of study.

REFERENCES

Abramson, N. (1963). *Information theory and coding.* New York: McGraw-Hill.

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution.* Berkeley, CA: University of California Press.

Berwick, R.C., & Weinberg, A.S. (1984). *The grammatical basis of linguistic performance: Language acquisition and use.* Cambridge, MA.: MIT Press.

Dorr, B.J. (1993). *Machine translation: A view from the lexicon.* Cambridge, MA: MIT Press.

Greenberg, J.H. (1966). Some universals of grammar with particular reference to the order of meaningful elements. In J.H. Greenberg (Ed.), *Universals of grammar.* Cambridge, MA: MIT Press.

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

Li, M., & Vitányi, P. (1997). *An introduction to Kolmogorov complexity and its applications* (2nd ed.). New York: Springer.

Lightfoot, D. (1991). *How to set parameters: Arguments from language change.* Cambridge, MA: MIT Press.

Morgan, J.L. (1986). *From simple input to complex grammar.* Cambridge, MA: MIT Press.

Nichols, J. (1992). *Linguistic diversity in space and time.* Chicago, IL: University of Chicago Press.

Perkins, R.D. (1992). *Deixis, grammar and culture.* Amsterdam: John Benjamins.

Pierce, L. (1993). *The ONLINE BIBLE user's guide.* Ontario, Canada.

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

Zipf, G.K. (1949). *Human behaviour and the principle of least effort.* New York: Hafner Publishing Company. (Reprinted 1965)

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

Ziv, J., & Lempel, A. (1978). Compression of individual sequences via variable rate coding. *IEEE Transactions on Information Theory, IT-24,* 530–536.