QUANTITATIVE ASSESSMENT OF TEXTUAL COMPLEXITY

Vincenzo Gervasi
Vincenzo Ambriola

When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science.

William Thomson, Lord Kelvin
Popular Lectures and Addresses (1891-1894)

0. *Introduction*

Consider the following statement:

> Mary had a baby.[1]

and compare it with Lord Kelvin's statement above. There is little doubt that most readers would describe Lord Kelvin's quote as *more complex* than the statement about Mary. Still, what *exactly* makes it more complex is unclear. The first statement is longer (with respect to several possible definitions of *length* or, more generally, of *size*), has a deeper syntactic structure, uses a more varied lexicon, and so on. But what relationship all these differences have with its complexity?

Textual complexity is indeed an elusive and multi-faceted quality, an attribute of texts of which we currently have only an unsatisfactory intuitive understanding. There is probably no hope to express such an attribute with a single "magic" number, as we do with simpler attributes like *length*.

A look backward to the history of our understanding of complex attributes of things, however, provides some encouragement. For many centuries, for example, we talked of *color* attributes only by referring to them with vague words (e.g., "blue"), by establishing imprecise analogies with colors found in nature (e.g., "peach"), or by creative use of adjectives and other specifications. Today, we can refer to various facets of "color" with great precision: to the hue as a frequency in the electromagnetic

[1] All sample texts not explicitly attributed are artificial.

spectrum, to the luminance as a multiple of a standard light emitter (the candle), to the saturation as a percentage, etc.

Our purpose in this paper is to propose some measures of a few facets of textual complexity. In particular, we will keep an engineering approach, and concentrate on quantitative measures that can be obtained through objective, repeatable methods, and possibly by using automatic tools and procedures. In the process, we will challenge our intuitive understanding of complexity, in order to define some *axioms* that will help us in our task.

1. *Measures and Metrics*

According to Fenton [FP97],

> **Measurement** is the process by which numbers or symbols are assigned to attributes of entities in the real world in such a way as to describe them according to clearly defined rules.

More precisely, measurement consists in mapping attributes of real-world entities to elements of some formally-defined set (often, a numeric set), and in mapping empirical relationships between the same entities to formal relationships between the corresponding elements of the formal set.

As an example, let us consider an apparently well-understood attribute like *length* of a text. We can define a simple mapping L between a text t and a natural number, meant to convey the essence of *length*, in this way:

$$L(t) = \text{number of words in } t$$

Moreover, we can map the intuitive relationship *longer-than* between texts to the formal relation *greater-than* (written $>$) between natural numbers. Thus, given two texts t and t', we can say that t is longer than t' if and only if $L(t) > L(t')$. Naturally, giving a precise definition does not spare us from possible counter-intuitive results. In fact, with the definitions above, the text

> Mary had a baby. She was fair and had blue eyes. I was in love with the little angel!

turns out to be exactly as long as

> Guaranteeing satisfactory results without appropriately substantiating unobvious assertions is inadmissible. Measurements demand thorough examination, formal verification, and intuitive acceptability.

A different definition of $L(t)$ (e.g., number of characters in t) might be more in line with our intuition in this case. It is not difficult to be convinced that the quality of a measure is strictly related to its ability to accurately reflect our intuition about the phenomenon that is being measured.

While in our example natural numbers were used as destination for the mapping, in many cases other sets may be more suited. For example, the attribute *style* of a text may be classified "high", "pompous", "familiar", and so on. As far as our intuition goes, it is pointless to define *style* as a number. This is not by chance: different attributes have different properties and need different *measurement scales*. There are five well-know classes of measurement scales, in increasing order of expressiveness:

- **Nominal scale.** Entities are simply collected in similarity classes, and no ordering exists between classes. Our definition of a measure for *style* was on a nominal scale: there is no ordering between "pompous" and "bureaucratic", and no magnitude is associated with the various classes.

- **Ordinal scale.** This scale adds an ordering relationship between the classes of the nominal scale. For example, the *understandability* attribute of a text could be measured on an ordinal scale as "low", "medium", "high", in increasing order. In order for the representation to be valid, the order established between the symbols "low", "medium" and "high" must be consistent with the intuitive ordering based on understandability between texts. There is still no magnitude associated with the measure.

- **Interval scale.** Interval scales add the capability to express the difference between two measures. Thus, if we define a measure of the *period* of a text based on the century the text was written, we can certainly say that a XV century text was written after a XIII century text, and — more precisely — that it was written two centuries later than the XIII century text. On an interval scale, addition and subtraction are legitimate, while multiplication and division are not: we cannot say that the period of a text is twice the period of another text.

- **Ratio scale.** These last two operations become legitimate on a ratio scale. These scales are characterized by the presence of a *zero element*, denoting the total absence of the attribute that is being measured. For example, our definition of *length* above is a ratio scale, since we have a zero element — the empty text, having length zero. Moreover, the measurement interval that starts at the zero element increases at fixed steps, called *units*. In our case, the units for L were single words.

- **Absolute scale.** The more precise scale is the absolute scale, that simply counts occurrences of something characterizing the attribute in the entity. Absolute scales are unique: there cannot be two different absolute scales for the same attributes. For example, our measure L is a ratio scale for *length* (we have seen that we could also have counted characters or lines instead of words), but it is an absolute scale for *number of words*.

It can be noted that each scale class includes all the properties of the preceding ones. So, using our measure for *length*, we can distinguish whether two texts are of equal length or not (nominal scale), we can say whether a text is longer or shorter than another text (ordinal scale), we can measure the difference in length between two texts (interval scale), and can say that the first text is exactly, say, 1.86 times as long as the second one (ratio scale).

With the conceptual tools that we presented above, collectively known as the *representational theory of measurement*, we are able to define precise and mathematically well-founded measures of simple attributes. In particular, we are interested in *direct* measures, that is, in attributes whose value can be determined by direct observation of the entity that is being measured. On the contrary, *derived* measures rely on relationships established by mathematics or by physical laws to obtain the value of the attributes. For example, the attribute *temperature* of a room[2] is usually measured indirectly by measuring directly the *length* of a column of mercury.

Unfortunately, in the case of very complex attributes — like *complexity* of a text — we cannot rely on precise relationships like those provided by physical laws. Instead, the so-called Factors-Criteria-Metrics (FCM for short) [MRW77] can be used.

In the FCM model, a quality that cannot be measured directly is instead estimated by using a number of *quality factors*, i.e., other, simpler attributes that have a strong positive correlation with the principal quality. Factors are usually defined aiming at minimum overlap, and thus at maximum orthogonality among them. This allows each factor to be controlled independently, and maximizes the efficiency of information collection. In addition, factors are supplemented by a set of support properties called *criteria*, which serve as indicators or predictors of factors. Criteria may be decomposed hierarchically in case they are not directly measurable properties. In that case, a derived measure can be assigned to a criterion by composing (by using some appropriate accumulation function) the measures of the leaves of its decomposition tree.

Finally, a *metric* is a direct measure of a criterion or sub-criterion, in the sense of our initial definition. In particular, metrics must be defined by specifying an attribute

[2]We are oversimplifying here; a more precise definition would be that of "temperature of the air at a certain point inside a room".

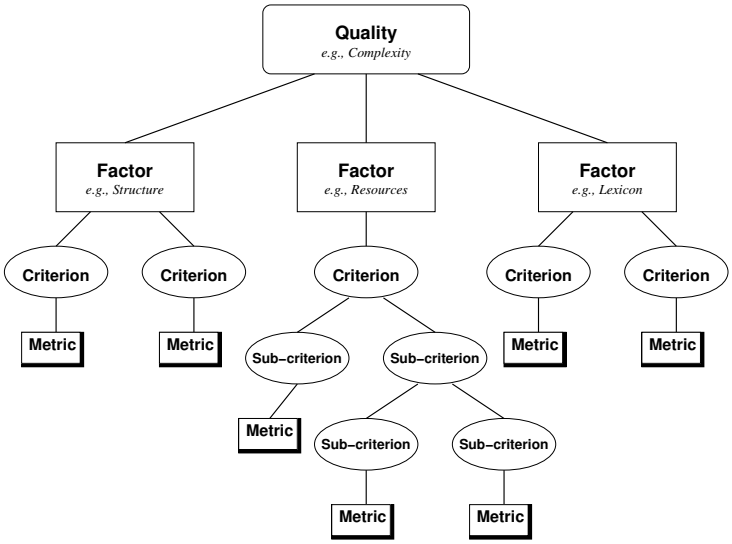


Figure 1: The Factors-Criteria-Metrics model.

of the entity and the exact procedure for measuring it. The overall structure of a FCM model is depicted in Figure 1.

For example, let us consider the complex attribute *size* of a text. We do not know how to measure *size* (and indeed, we do not even have a precise definition of what we mean by *size*), but we know that *length* influences *size*, and thus, *length* can be considered to be a factor for *size* (together with other factors, such as *size of lexicon*). In turn, *number of words* is a criterion for *length* (together with other criteria, such as *average word length*). To measure *number of words* we define a precise metric, detailing what is a "word", counting rules, how to treat special cases like hyphenated words, and other details that guarantee objectivity and repeatability of the measure.

In this work, we do not try to define a complete FCM model for *complexity*. Rather, we discuss a selection of possible criteria and metrics, with a particular emphasis on the possibility of automatically obtaining the measures from the text.

2. *Axioms of complexity*

In the previous section we have seen that any measure is only as good as its capability to reflect our intuitive understanding of the phenomenon that is being measured. The

same principle holds for the metrics of the FCM model. But what is our "intuitive understanding" of textual complexity? In this section, we try to clarify this point by considering a number of *axioms* for the attribute *complexity*. It is important to remark that we are not trying to give an ultimate definition of complexity through these axioms — in other words, we are not saying which of the proposed properties are really axioms that are held true with no need for proof. However, concrete definitions of metrics for *complexity* will exhibit certain properties based on those among the axioms (below) that are satisfied by the metric.

Notation. In the following, we indicate with t, t' etc. an arbitrary text, and with $\mathcal{C}$ a measure for *complexity*. We use the standard set-theoretic and logic notation applied to texts: $t \subseteq t'$ means that text t' includes text t, and $t \cap t'$ indicates the common part of text between t and t'. To indicate the text obtained by appending t' at the end of t, we write $t \cdot t'$. We use the symbol $\emptyset$ to denote the empty text.

Axiom 0 (Zero). We first investigate the complexity of the empty text. If complexity has a minimum, then the empty text is a good candidate to represent the less complex text possible. In formal terms,

$$\forall t \neq \emptyset, \mathcal{C}(t) > \mathcal{C}(\emptyset)$$

However, even this basic property is not to be taken for granted. What is the complexity of silence? In terms of size or structure, we could assign the lowest possible complexity to the empty text, but some measure could consider "silence" to be the most difficult text to decode, and possibly assign the highest possible complexity to it.

Axiom 1 (Monotonicity). A second interesting point is whether complexity is *monotonic*, i.e., if by adding something to a text, its complexity does necessarily increase. In formal terms, we want to determine if, given a certain measure $\mathcal{C}$,

$$\forall t, t',\ t \subseteq t' \Rightarrow \mathcal{C}(t) \leq \mathcal{C}(t')$$

is true or not. Monotonicity is an important characteristic, but the answer is not self-evident. For example, t' could contain an explanation for something that is stated in t, and in this case the complete text t' could well be considered less complex than its unexplained fragment t. On the other hand, our understanding of complexity could tend to correlate the *complexity* of a text with its *size* (longer texts are more complex than shorter texts), and in this case monotonicity would clearly hold.

Monotonic metrics tend to be on ratio or absolute scales. In fact, the definition above lends itself naturally to the convention $\mathcal{C}(\emptyset) = 0$, with any other text having higher complexity than the empty text.

Axiom 2 (Compositionality). Can the complexity of a text be determined exclusively from the complexity of its constituent parts? If this is the case, such a measure is said to be *compositional*. Formally, given a function T to compose two texts into a larger text, and a function M to compose two measures into a single measure, compositionality can be expressed as follows:

$$\forall t, t',\ \mathcal{C}(T(t, t')) = M(\mathcal{C}(t), \mathcal{C}(t'))$$

For example, if we take the append operator $\cdot$ as T and algebraic sum $+$ as M, we can rewrite the property above as

$$\forall t, t',\ \mathcal{C}(t \cdot t') = \mathcal{C}(t) + \mathcal{C}(t')$$

meaning that if we want to measure the complexity of a two-parts text, we can measure the complexity of the first part, then the complexity of the second part, and add the two together (the process extends in the obvious way to more than two parts).

Compositional measures have great practical and theoretical advantages. Such measures can be defined by assigning values to some basic cases, and then giving rules that relate the way in which texts are built from the basic cases, and the way in which complexity grows when texts are composed. As a pleasant side effect, compositionality guarantees that any observation made on small examples is also applicable to texts of arbitrary size.

Axiom 3 (Classifiability). It would be very surprising if each and every text had its own complexity, different from that of any other text. Such a property would render text classification impossible: each text would constitute a class on its own. On the contrary, we can expect that, apart from some particular case, given an arbitrary text t we can find a different text t' that has the same complexity. In formal terms, this property can be written

$$\forall t \neq \emptyset, \exists t' \neq t \text{ such that } \mathcal{C}(t) = \mathcal{C}(t')$$

Metrics that satisfy classifiability permit the creation of classes of text of equal complexity. If complexity is measured on at least an ordinal scale, these classes can also be ordered — a capital property for practical applications.

Axiom 4 (Structurality). Complexity can be assigned to some aspect of a text rather than to the entire text as a whole. In particular, it is not unreasonable to assume that complexity resides in the structure (of some kind, e.g. syntactic structure) of a text rather than on the particular concrete elements (e.g., particular words) of the text. Formally, given an abstraction function S that extracts some structural aspect of a text, we can define structurality as

$$\forall t, t',\ S(t) = S(t') \Rightarrow \mathcal{C}(t) = \mathcal{C}(t')$$

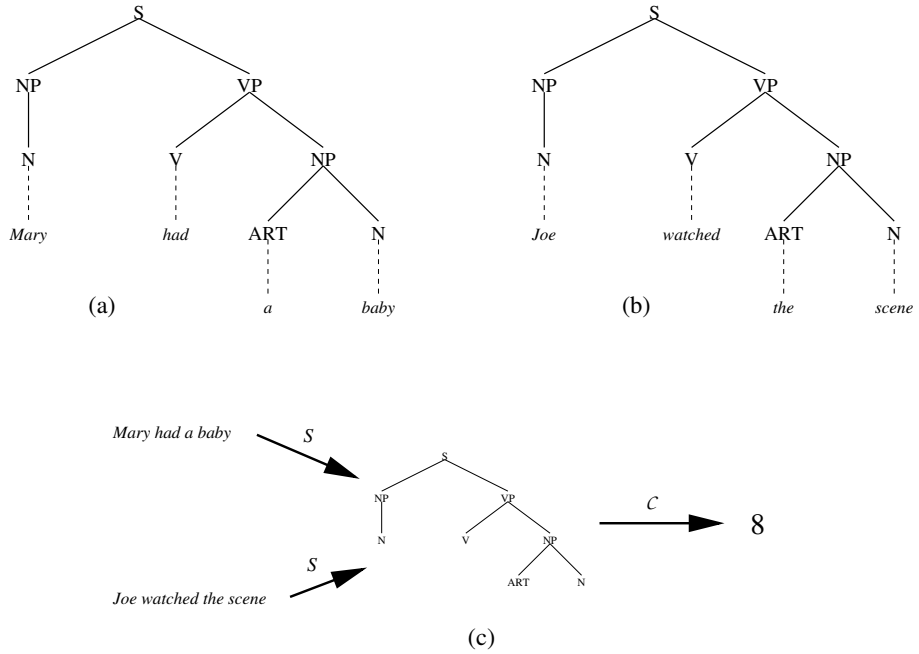


Figure 2: A sample structural metric. (a) and (b), syntactic parse trees for two statements. (c), how a structural metric measures only attributes of the structure of the texts.

As a typical example, if S is a function to extract the syntactic structure of a text, a metric obeying structurality would declare that two texts with the same syntactic structure have the same complexity — regardless of the particular nouns, verbs, adjectives, etc. that appear in the texts. Figure 2 illustrates this property: here we consider syntactic structure (as obtained by a standard parser for English) as our S. The two texts "Mary had a baby" (t) and "Joe watched the scene" (t') have the same syntactic structure, hence the measure $\mathcal{C}$ (in the example, a simple count of the nodes in the parse tree) returns the same value for both texts.

Axiom 5 (Unstructurality). The property that is dual to structurality is also reasonable, and assumes the same formal expression. Given a flattening function F, i.e., a function that discards the structure from a text, we can define unstructurality as

$$\forall t, t',\ F(t) = F(t') \Rightarrow \mathcal{C}(t) = \mathcal{C}(t')$$

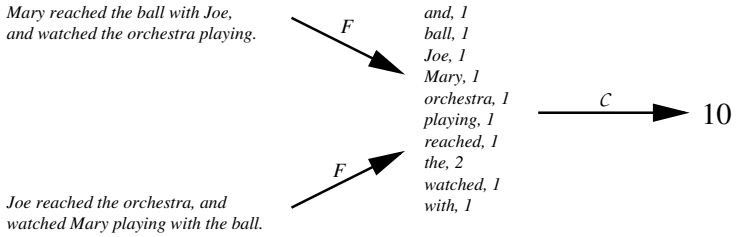


Figure 3: A sample unstructural metric.

For example, F could be a function that, given a text, returns simply the set of the words that appear in the text, possibly with the respective multiplicity. Such a function would flatten-out the syntactic structure; this property states that texts using the same words have the same complexity. Figure 3 illustrates this definition.

It is important to remark that structurality and unstructurality are indeed the same property in formal terms; only our interpretation of the meaning of the abstraction function used (S to extract structure, or F to flatten it out) determines the kind of property that we are using.

Axiom 6 (Order independence). As a particularly relevant case of unstructurality, we consider the dependence of the metric for complexity on the order in which the text is presented. Shall a question, followed by an answer, have the same complexity as the same answer, followed by the question? In formal terms,

$$\forall t, t',\ \mathcal{C}(t \cdot t') = \mathcal{C}(t' \cdot t)$$

states that the metric $\mathcal{C}$ is independent of the order in which the constituent parts of a text appear.

This list of axioms is not intended to be exhaustive. More axioms could be defined, to express the desired behavior of *complexity* (better reflecting our intuitive understanding of that attribute) or desired computational properties of its measures. However, the list above already constitutes a good characterization of *complexity*, and any desired addition would not affect the observations made in the following.

3. *A smorgasbord of complexity metrics*

In this section we present a number of metrics (in the sense of the FCM model of Section 1) that have in common their amenability to automatic extraction and processing.

While all the metrics presented have some correlation with intuitive *complexity*, we do not prescribe how these metrics should be used to estimate criteria and factors of *complexity*. This task is better left to a discussion of the ontological nature of *complexity*. Here we take the pragmatic approach of providing estimators: we concentrate on the process of obtaining estimators, and leave the actual estimation to further work. However, for presentation purposes, metrics are collected in classes, that could hint at which factors the metrics relate to.

3.1. *Readability metrics*

A wide array of metrics has been developed in the second half of the last century for *readability*, from the venerable Flesch Readability Index, to many recent proposals (see Masi in this volume). It is sensible to assume that *readability* is a factor of *complexity*, under the hypothesis that complex texts are more difficult to read. Readability metrics are thus relevant for our purposes.

Table 1, compiled with the help of [MP82], collects a number of early proposals for *readability* metrics. All the metrics combine various features of texts in different ways and with different parameters. However, if we look at these formulas in the light of our axioms, we discover that they measure very similar attributes. This is an indication that the underlying intuition about *readability* is shared among all the authors.

All the metrics are *density-like* measures; they do not depend on the size of the text. In fact, all variable terms are divided by W (number of words in the text) or by T (number of sentences in the text), both of which are estimators for *size*. Since all the other parameters are also correlated with *size*, and they only appear in fractions, dimensional analysis ensures us that the formulas produce adimensional results.

As a consequence, readability measures do not satisfy our axioms Zero[3] and Monotonicity. All these metrics for *readability* are not compositional, due to the presence of density factors (S/W, W/T, M/W, etc.). We can easily prove it by contradiction.[4] Let us assume that $\mathcal{C}(t) = S/W$ is a measure for *readability*, and that the textual composition function T does not discard part of the text. Then by Axiom 2 we have:

$$\frac{S+S'}{W+W'} = \mathcal{C}(T(t,t')) = M(\mathcal{C}(t),\mathcal{C}(t')) = M(\frac{S}{W},\frac{S'}{W'})$$

but we also have that

$$M(\frac{S}{W},\frac{S'}{W'}) = M(\frac{2S}{2W},\frac{S'}{W'}) = \frac{2S+S}{2W+W}$$

[3] In effect, none of the measures can be computed for an empty text, since in that case $W = T = 0$.

[4] We prove it for a single density factor; the proof can be applied to complete formulas as well. Numeric factors are non-essential to the proof.

Author	Formula	Reference
Flesch	$R = 206.835 - 84.6\frac{S}{W} - 1.015\frac{W}{T}$	[Fle48]
Farr, Jenkins, Paterson	$R = -31.517 + 159.9\frac{M}{W} - 1.015\frac{W}{T}$	[FJP51]
Dale, Chall	$G = 19.4265 - 15.79\frac{D}{W} + 0.0496\frac{W}{T}$	[DC48]
Powers, Sumner, Kearl	$G = 14.8172 - 11.55\frac{D}{W} + 0.0596\frac{W}{T}$	[PSK58]
Holquist	$G = 14.862 - 11.42\frac{D}{W} + 0.0512\frac{W}{T}$	[Hol68]
Gunning	$G = 3.0680 + 9.84\frac{P}{W} + 0.0877\frac{W}{T}$	[Gun52]
Coleman	$R = -37.95 + 116.0\frac{M}{W} + 148.0\frac{T}{W}$	[Col65]
McLaughlin	$G' = 3.1291 + 5.7127\sqrt{\frac{P}{T}}$	[McL69]

Measures

R = Readability index (0–100)
G = Grade (0-12) to answer 50% of the questions about a text
G' = Grade (0-12) to answer 100% of the questions about a text

Parameters

W = total number of words
T = total number of sentences
L = total number of letters
V = total number of vowels
D = total number of words in the Dale Long List [DC48]
S = total number of syllables
M = total number of monosyllabic words
P = total number of words with 3 syllables or more

Table 1: A collection of metrics for *readability* proposed in the literature.

thus leading to the contradiction:

$$\frac{S + S'}{W + W'} = \frac{2S + S'}{2W + W'}$$

Since they are based exclusively on a count of the number of words, sentences, syllables, etc., all the metrics are clearly unstructural and order-independent. In fact, according to these metrics, the text

> Mary had a baby. She was fair and had blue eyes. I was in love with the little angel!

has exactly the same *readability* as

> A and angel, baby blue eyes! Fair had had I in little love Mary. She the was was with.

It appears thus that *readability* is only a marginally useful indicator for *complexity*, covering almost exclusively the unstructural aspects.

3.2. *Information content metrics*

Another classical field of studies that is relevant for our purposes is Information Theory. Originated from the seminal works of Nyquist [Nyq24, Nyq28] and Hartley [Har28], and brought to maturity by Shannon [Sha48], Information Theory deals with the *information content* of a signal. The hypothesis here is that texts conveying more information are more complex, or, in terms of resources, that more information in a text requires more memory and more time to the reader, thus increasing the perceived complexity.

A text is modeled as an *ergodic source*; in other words, a text is seen as a series of *symbols* (letters, words, etc.), where each symbol appears with a certain probability (that can depend on the appearance of previous symbols). At each step, the source has a certain degree of *freedom* about the next symbol to be produced; from the point of view of the receiver (the reader of the text), this freedom corresponds to a certain degree of *uncertainty* about the next symbol that will be read.

The amount of information that is carried by a text is directly related (and is actually a measure of) this freedom/uncertainty. In fact, if we imagine a source that has no choice — say, always producing a symbol α —, the reader does not gain any knowledge by examining the message: she would know beforehand what the "content" is. After all, nobody wants to read a page of αs.

If all symbols appear with the same probability, there is no way for the reader to guess the next symbol without actually looking at the message. For example, if we

are tossing a coin and writing down the sequence of outcomes, the person reading the text will not know what to expect next after having read only part of the text. In this case, the information content of the text is maximal: the coin has maximum freedom, and the reader maximum uncertainty.

English as a language has much less freedom, since it exhibits an extremely rich structure on all levels, from letter sequences to argumentative structure. At the lower level, certain letters are more frequent than others in English. Our measures on a corpus of 17 million characters (contemporary technical English), show that the most frequent letter, 'e', has a frequency of 11.46%, while the less frequent one, 'z', has a frequency of 0.14%. Thus, in our sense, a 'z' in the text conveys more information than an 'e'. If we examine whole words, we find that the most frequent word, 'the', has a frequency of 5.13%, while other words like 'zealously' (among many others) have a frequency as low as 0.000000448995%. Thus, an occurrence of 'zealously' is more characterizing, carries more information, and presumably adds more complexity to the text, than an occurrence of 'the'.

The classical measure of information content is *entropy*, that is defined formally as

$$H = -\sum_{s \in A} \mathbb{P}(s) \log_2 \mathbb{P}(s)$$

where A is the set of all the symbols (letters, words, etc.), and $\mathbb{P}(s)$ is the probability of occurrence of symbol s. The measure H that is obtained, expressed in *bits* (short for Binary digIT), is on an absolute scale: the measure counts the minimum number of bits per symbol that are needed to encode the message produced by the source, i.e. the text.

Naturally, in English the probability of occurrence of a certain symbol depends on what has already appeared. In certain cases the dependence can be very strong: a search in a dictionary finds 420 words containing the sequence THE, but only one (earthquake) containing the sequence THQ. The interpretation of this fact is twofold. First, this means that after having seen a 't' and a 'h', the probability of seeing an 'e' is very high (thus 'e' brings little new information), while the probability of seeing a 'q' is very low (thus 'q' brings much information). Second, and confirming this last point, we know that seeing a 'q' after 'th' uniquely identifies the word "earthquake": all the preceding and following letters bring no new information at all.

The same reasoning applies to sequences of words, of parts-of-speech, of semantic categories, etc. In the case of sequences, entropy is defined as follows:

$$H_n = -\sum_{B \in A^n, s \in A} \mathbb{P}(B) \mathbb{P}_B(s) \log_2 \mathbb{P}_B(s)$$

sample	size	L-H	W-H_0	W-H_1	W-H_2
C.T.E.	2227198	4.233	10.778	6.395	2.670
Flatland [Abb84]	33587	4.158	9.173	4.506	1.154
When you can measure...	62	4.179	4.912	0.850	0.113
Guaranteeing satisfactory results...	19	4.045	4.248	0	0
Mary had a baby...	19	3.909	4.037	0.222	0

Table 2: Measures of entropy on various texts.

where A^n denotes the set of all the sequences of symbols from A of length n, $\mathbb{P}(B)$ denotes the probability of occurrence of the sequence B, and $\mathbb{P}_B(s)$ denotes the probability of occurrence of a symbol s after a sequence of symbols B. Progressively larger values of n provide more precise measurement of the way in which inter-symbol relationships of increasing distance influence the information content.

Table 2 presents the results of our entropy measures on various texts: the large Contemporary Technical English (C.T.E.) corpus that we used above, a moderate size narrative, and three of our sample texts. As can be observed, letter-based entropy L-H is substantially independent of *size* (that is measured as "number of words" in Table 2, where a word is any sequence of symbols surrounded by white space). Moreover, it appears that the technical language used in the C.T.E., and the 19^{th}-century prose in Flatland and in Lord Kelvin's statement produce slightly higher measures for L-H than the other simpler texts.

The difference is even more striking when we consider word-based entropy, W-H_0 (H_0 is equivalent to H, according to the definition of H_n above). In this case, the C.T.E. and Flatland have much higher information content than any of the sample texts; among the samples, Lord Kelvin's statement scores more complex, and our "Mary had a baby" sample scores less complex — that is well in accord with our intuition.

Unfortunately, the sample texts are too short to allow significative measurement of word sequence-based entropy — the values tend to 0 since sequences of 2 or 3 words tend to be unique in such short texts. The values for C.T.E. and Flatland are once again higher, as we would expect. In particular, for C.T.E., a W-H_1 value of 6.395 means that, on average, given a single word, there are slightly more than 84 ($2^{6.395}$) words that can follow the word given. There is thus a substantial uncertainty about what will actually follow after a certain word. Even when two words in sequence are given, there are still — on average — more than 6 words among which the third can be chosen, according to the value for W-H_2. We really need to read the entire text to know what the text is saying — an indication that the text has indeed substantial *complexity*.

Let us now consider which ones among our axioms are satisfied by information content measures (with special reference to entropy-based measures). First, axiom Zero is satisfied, with $\mathcal{C}(\emptyset) = 0$. In fact, for an empty text we have $A = A^n = \emptyset$, and thus the result of the sum is 0. Entropy is not monotonic; the way in which the measure changes depends on the nature of the text that is added (and, in particular, on the way in which the new text changes the frequency of each symbol). For example, taking $\mathcal{C}(s)$ equal to H for a source s (with independent probabilities), we have that

$$\mathcal{C}(\text{`a b a'}) = 0.918 \qquad \leq \qquad \mathcal{C}(\text{`a b'}) = 1 \qquad \leq \qquad \mathcal{C}(\text{`a b c'}) = 1.585$$

As can be seen, depending on what we add after 'a b', entropy can either increase or decrease. Moreover, entropy is not compositional: again, the way in which frequencies change when composing texts is not fixed. Let us take as our text composition operator T the append operation, and assume that there exists a function M to compose the measures, as per axiom 2. Then, we have

$$\begin{aligned}
2 &= \mathcal{C}(\text{`a b c d'}) \\
&= \mathcal{C}(T(\text{`a b'}, \text{`c d'})) \\
&= M(\mathcal{C}(\text{`a b'}), \mathcal{C}(\text{`c d'})) \\
&= M(1, 1) \\
&= M(\mathcal{C}(\text{`a b'}), \mathcal{C}(\text{`a b'})) \\
&= \mathcal{C}(T(\text{`a b'}, \text{`c d'})) \\
&= \mathcal{C}(\text{`a b a b'}) \\
&= 1
\end{aligned}$$

The contradiction $2 = 1$ proves that no such M can exist, and thus that entropy-based measures are not compositional.

We have already seen in the examples that entropy satisfies our classifiability axiom. Regarding structurality and unstructurality, the nature of a measure depends on the particular kind of symbols that is being analyzed. If we take letters or words as symbols, the measure is clearly unstructural:

$$\mathcal{C}(\text{`Mary had a baby'}) = 2 = \mathcal{C}(\text{`A baby had Mary'})$$

However, if we consider higher-order structures, and especially if we take H_n for $n > 0$, the measure tends to become structural. For example, we can decide to consider parts-of-speech as symbols, instead of words. The text "Mary had a baby" would be mapped (by the abstraction function S of axiom 4[5]) to the sequence of tags NN VB

[5]This particular abstraction function is implemented in a number of popular *parts-of-speech taggers*.

DT NN (noun, verb, determiner, noun). The text "Joe watched the scene" produces the same sequence of parts-of-speech, thus

$$S(\text{'Mary had a baby'}) = S(\text{'Joe watched the scene'})$$
$$\Rightarrow \mathcal{C}(\text{'Mary had a baby'}) = \mathcal{C}(\text{'Joe watched the scene'})$$

that is exactly our definition of structurality. In the cases in which $\mathcal{C}$ is unstructural, order-independence is also satisfied, for $n = 0$. It is never satisfied for $n \geq 1$, since in this case the ordering of symbols affects the probability of occurrence of blocks that are used by the measure.

In conclusion, metrics based on information content appear to measure relevant facets of *complexity*, can be easily computed by using the definition of entropy, have a good measurement scale, and can be used to measure structural or unstructural aspects, according to necessity, by using different kinds of symbols and block lengths.

3.3. *Structural metrics*

In the previous section we have seen how entropy can be used to measure some attribute of the structure of a text. However, while a sequence of symbols is a close model for a text (taken as a sequence of letters or words), it is a poor model for more complex structures: and every text of practical interest has many of these structures underlying the surface "sequence of words".

A large family of metrics can be obtained by explicitly identifying and building such structures, and by measuring their attributes through direct metrics.

The obvious candidate for a structural metric is some measure concerning the syntactic structure of a statement. A *parse tree* for a statement can be obtained by using any of a number of grammars, formalisms, and tools that are widely available (see [All95] for a survey of the field). For our purposes, it is not particularly relevant which tools we choose: rather, we are mainly interested in finding an automatic way of building parse trees, in order to guarantee consistence and repeatability of the measures. In the following, we use our domain-based parser CICO [Ger01], with a simple ad-hoc grammar for English, but other parsers could be used as well.

CICO is a parser based on fuzzy matching of textual *templates* to fragments of a text. The matching is based on a set of conjunctive types: each term in the text has associated a set of tags, that can denote either grammatical properties (for example: noun, verb, singular, superlative, etc.) or semantic, domain-specific properties (for example: human being, event, time span, etc.). We assume in this section that only

We used the publicly available TreeTagger [Sch94] in our tests. The tag codes we use are inspired by those in [MSM93].

grammatical tags are assigned to terms, by using a parts-of-speech tagger like the one we used in the previous section. CICO's rules assume the form

$$\langle \textit{Model}, \textit{Action}, \textit{Substitution} \rangle$$

where the *Model* is a template for a syntactical structure, in which variable parts are matched according to their tags; the *Action* is a semantic annotation (in practice, instructions on which kind of node to add to the parsing tree that is being constructed), and the *Substitution* specifies how the fragment should be treated for the purpose of further analysis. For example, the rule

$$\langle \texttt{ det/DT/0 adj/JJ/0 n/NN, NP \$det \$adj \$n, \$ID/NP } \rangle$$

specifies that upon matching an optional[6] determiner, followed by an optional adjective, followed by a noun, the parser should emit a node of type `NP`, with children nodes corresponding to the various parts of the template, and continue the parsing substituting the `NP` node for the whole fragment. Due to the fuzziness of the parser, a matching can be imperfect in several ways (constituents can be out of order, extra or missing terms are allowed, etc.). Moreover, since the parser employs a heuristic backtracking strategy, several rules can match the same fragment, and the parser will choose the one that maximizes the overall score of the parse tree. Both these features have a positive impact on the robustness of the parsing, and allow reasonably compact grammars to be used.

Once a set of parse trees is obtained for a text (see for example Table 3), various simple measures can be obtained. A simple count of the number of nodes in the trees provides an estimation of the complexity of the syntactic structures — clearly a factor for *complexity* of the whole text. The same attribute can be measured in slightly different ways, e.g. by measuring the depth of the deepest leaf in the tree, or by weighting the number of leaves according to the depth. Measures for the three sentences in Table 3 are reported in Table 4. The type of each node can also be taken into account: prepositional phrases may conceivably add more complexity than noun phrases.

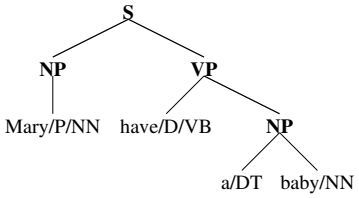
Since metrics on syntactic structures are defined at the level of single sentences, we are free to compose the measures in such a way as to preserve desirable properties of the metric. We can define the measure $\mathcal{C}$ for an entire text t as the sum of the measures for each sentence s in the text, as in

$$\mathcal{C}(t) = \sum_{s \in t} \mathcal{C}(s)$$

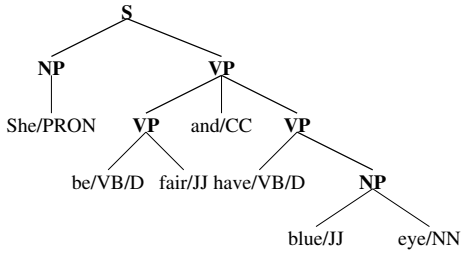
[6]Optionality is denoted by the `/0` tag in the template.

Model	Action	Substitution
det/DT/0 adj/JJ/0 n/NN	NP $det $adj $n	$ID/NP
p/PRON	NP $p	$ID/NP
prep/IN n/NP	PP $prep $n	$ID/PP
p1/PP p2/PP	PP $p1 $p2	$ID/PP
v/VB n/NP/0 p/PP/0	VP $v $n $p	$ID/VP
BE/VB b/JJ	VP BE $b	$ID/VP
v1/VP c/CC v2/VP	VP $v1 $c $v2	$ID/VP
n/NP v/VP	S $n $v	$ID/S

Mary had a baby.



She was fair and had blue eyes.



I was in love with the little angel!

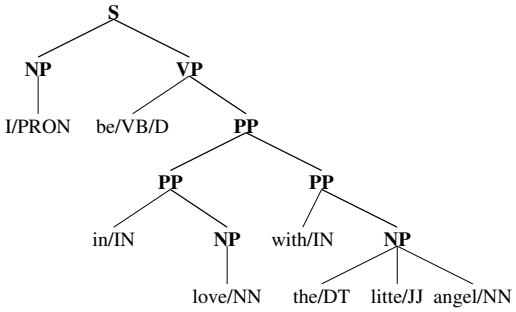


Table 3: Some parsing rules, and their application to a sample text.

Text	$\|L\|$	$\|N\|$	$\max_{n\in L} d_n$	$\sum_{n\in L} \log_2 d_n$
Mary had a baby.	4	8	3	5.17
She was fair and had blue eyes.	7	13	4	10.75
I was in love with the little angel!	8	16	5	15.29

Table 4: Some measure of structural complexity on the sample statements of Table 3. N is the set of all the nodes in the parse tree, $L \subset N$ is the set of the leaf nodes, d_n is the depth of node n.

where by $s \in t$ we indicate that s ranges over the *sequence* of statements t. In this case, the measure satisfies many of our axioms: in fact,

- $\mathcal{C}(t)$ satisfies axiom Zero, since the sum of no terms is 0;
- $\mathcal{C}(t)$ is monotonic, since $t \subseteq t' \Rightarrow \exists t'' : t' = t \cup t'' \wedge t' \cap t'' = \emptyset$ and thus $\mathcal{C}(t') = \mathcal{C}(t) + \mathcal{C}(t'')$. By definition of $\mathcal{C}$, $\forall t, \mathcal{C}(t) \geq 0$; it follows that $\mathcal{C}(t') \geq \mathcal{C}(t)$, thus proving monotonicity.
- $\mathcal{C}(t)$ is compositional, taking $T(t, t') = t \cdot t'$ and $M(a, b) = a + b$. In fact,

$$\begin{aligned}
\mathcal{C}(T(t,t')) &= \mathcal{C}(t \cdot t') \\
&= \sum_{s \in t \cdot t'} \mathcal{C}(s) \\
&= \sum_{s \in t} \mathcal{C}(s) + \sum_{s \in t'} \mathcal{C}(s) \\
&= M(\sum_{s \in t} \mathcal{C}(s), \sum_{s \in t'} \mathcal{C}(s)) \\
&= M(\mathcal{C}(t), \mathcal{C}(t'))
\end{aligned}$$

  thus proving compositionality — in the case in which composition is performed on the level of full sentences. Similar but slightly more complex proof, that we omit here, shows that compositionality holds also for T functions that compose parts of sentences, as long as those Ts preserve syntactic correctness.
- $\mathcal{C}(t)$ satisfies the Classifiability axiom. In fact, given a text t, it is sufficient to change any of the words in t with another from the same syntactic category to obtain a text $t' \neq t$, that has the same syntactic structure and thus the same value under $\mathcal{C}$.

- $\mathcal{C}(t)$ is structural, as all the parameters in the various alternative definitions depend (only) on the parse trees.
- Finally, $\mathcal{C}(t)$ is order-independent on the level of whole sentences, since, for all t,t',

$$\mathcal{C}(t \cdot t') = \sum_{s \in t \cdot t'} \mathcal{C}(s) = \sum_{s \in t} \mathcal{C}(s) + \sum_{s \in t'} \mathcal{C}(s) = \sum_{s \in t' \cdot t} \mathcal{C}(s) = \mathcal{C}(t' \cdot t)$$

  On the other hand, the measure is not order-independent on the level of sentence constituents, since in this case a change of order may in general change the syntactic structure of the sentence.

The family of measures that we have so defined have a clear correlation with the *length* attribute of a text, since we are summing the measures across all the sentences of a text. In some application, e.g. when the metrics are used to evaluate the writing style of a text, this is not desirable. In these cases, we can define a new measure, that we will call *complexity density*, by dividing the *complexity* measure by a measure of *length*, for example *number of sentences*. *Complexity density* of a text can also be seen as the complexity of the average sentence in the text, that explains why it can be used as an indicator for *style*. Unfortunately, as happens with other density-like measures, a number of our axioms are no longer satisfied; the flexibility of the measure is thus reduced.

Naturally, ambiguity can prevent the identification of a single, certain parse tree for a sentence (for example, in the well-known case of prepositional phrase attachment). In turn, multiple possible parse trees lead to multiple different measures for the same sentence, making it impossible to compute a metric. This is not as bad as it may seem at first sight. In fact, we can well assume that this ambiguity adds some *complexity* of its own, that can be measured by counting the number of possible different parse trees, or by taking the highest complexity score among all the possible trees for the summation. In order to maintain a close correspondence with our intuition for *complexity*, whichever measuring rule is chosen for these cases, two properties should be maintained: (i) an ambiguous sentence should not measure less complex than an equivalent non-ambiguous one, and (ii) the measuring rule should not change the set of axioms that are already satisfied by the metric.

In our case, taking the maximum among the measures for all the alternative parse trees for a sentence satisfies both properties, and is an efficient and automatic[7] way of facing, if not solving, the ambiguity problem.

[7]The CICO parser can be instructed to output all the admissible parse trees for a sentence, instead of only the best-scoring one.

Although we concentrated on syntactic structures, these are not the only structures underlying language. Most other structures are organized in a tree fashion, and are amenable to the same kind of analysis. As an example, we can apply similar principles to the argumentative structure of a text. In related studies, we investigated how the rethorical structure of a sufficiently-marked text can be automatically retrieved by using techniques similar to the ones we used above for syntactic parsing. Rules like

```
⟨ x/THESIS IN FACT y/FACTS, PROBATION $x $y, $ID/PROBATION ⟩
```

allow the construction of graphs representing the argumentative structure of a text, on which measures in the style of the ones we presented above can be defined. See Figure 4 for an example of these techniques.

There is little doubt that a text with a complex syntax does not necessarily have a complex argumentative structure (although typically there is some correlation in sound texts), and vice versa. This is a clear indication that the respective metrics measure different facets of *complexity*. The same holds for other structures: from alliterative to dialogic structure, a huge number of structural phenomena can be measured.

3.4. *Semantic metrics*

Consider our old favorite,

> Mary had a baby. She was fair and had blue eyes. I was in love with the little angel!

and compare it with the following text:

> Malonylurea is a compound. It is crystalline and has ortho-pyramidal structure. It combines with ergotamine in an endothermic reaction.

Despite the difference of their subjects, these two texts are very similar: their syntactic structures are identical (see Table 3), they map to the same sequence of parts-of-speech (NN VB DT NN PP VB JJ CC VB JJ NN PP VB IN NN IN DT JJ NN), and even have the same word-based entropy (4.0374). Yet, many would rate the second text more complex than the first. Clearly, there is some facet of *complexity* that depends on *what* is said rather than on *how* it is said; in other words, part of *complexity* depends on semantics.

The exact nature of this dependence, however, is not easily defined. In fact, a text may seem complex to someone unfamiliar with the domain, but simple to someone familiar with it. A chemist would easily recognize our second text as referring to barbiturate drugs, and would probably spot a few blatant errors in the description of

A hyperhygienist mom has set a bucket next to a stainless-steel basin similar to the kind used for surgical instruments. Inside of it, immersed in a disinfectant solution, are a feeding bottle, a spoon, a small dish, and a teething ring for the baby to chew on. Mommy is nursing her hyperprotected young one, selecting the instruments with a pair of tongs. The latest do-it-yourself childcare vogue is taking hold among young families.
The whole affair would be edifying if it wasn't turning comical. Indeed mom hasn't got a pair of surgeon's gloves ready at hand before grandma can cuddle the tiny tot, and neither does she have a nurse's cap, lest the baby should pull at grandma's unsterilized hair. Further, grandma won't spray any disinfectant on the stroller belts, which the child cravingly sucks on regardless of all those who've laid their dirty hands all over the stroller. Lastly, she'll ignore that the stroller seat is nearly flush against the ground, and that the child will thereby inhale the blanket of exhaust fumes lying stagnant half a meter above the asphalt.
Better safe than sorry, you might say. But not necessarily so. My generation was always running around with dirty hands because we used to play in the streets witness the Tour of Italy, waymarked using chalk sticks, and the orange-soda caps sporting tiny pictures of Guerra and Binda. We'd stick two filthy fingers in our mouths to whistle. And we did wash our hands before eating, but then in the event that a piece of macaroni should fall off the table, we were taught to pick it up and blow on it to make it perfectly eatable again. If we peeled a knee by accident, we'd rinse it under an art-nouveau iron fountain. If the doctor pushed a spoon in our throats, he'd then wipe it clean with a handkerchief and the spoon was ready for another set of tonsils. The only implement he'd boil on a burner was the antitetanic syringe, but we were bent more on defying tetanus than getting pricked with a needle. And yet this untamed sort of lifestyle was antidotal for us: germs got around at large, and we feared our own antibodies. Today's viruses, on the other hand, go hand in hand with the toddlers of hygienist moms. I know one such mom who had her child skip the crawling phase altogether. (That's the phase in which children will grope around staking out their space by moving on all fours). The little kid had never touched the ground with his tiny hands, so when he went to the kindergarten he stuck his hands just about everywhere, then he sucked on his thumb and got all the neighbourhood germs.
And yet hygienist moms won't back down: they're breeding offspring that ought to live with gloves and respirators.

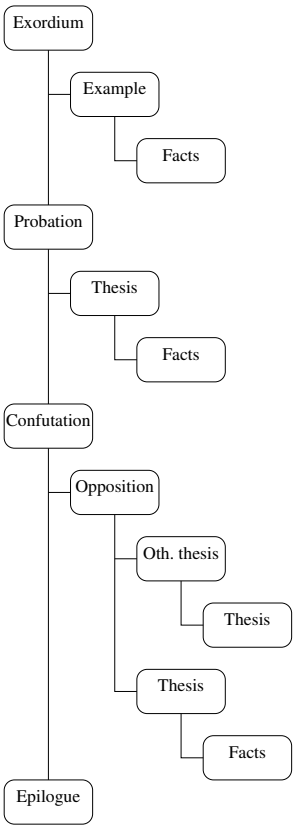


Figure 4: Argumentative structure for a sample text, as extracted by CICO with a rule set based on rethoric roles and constructions.

malonylurea — a feat impossible to the authors. Thus, is this kind of complexity objective? Probably not, but we can still measure it in reference to the knowledge body of a *standard reader*. Notice that we do not need an *average* reader; all we need is a reader that is perfectly objective, repeatable, immune to fatigue, boredom, or distraction, and a few other characteristics that make finding such a reader a hard goal.

As a more convenient approximation, we will resort to a well-known semantic resource, the WordNet database [Fel98] developed at the Princeton University's Cognitive Science Laboratory. In WordNet, English nouns, verbs, adverbs, and adjectives are organized into synonym sets (*synsets* for short), each representing one underlying lexical concept. Different relations link the synonym sets: among them, synonymy, antonymy, hyperonymy, hiponymy, holonymy, meronymy, etc. Figure 5 shows a (minuscule) fragment of the WordNet network surrounding *Malonylurea*.

Such a rich structure offers numerous measurement opportunities. For example, if we hypothesize that using a very specialized language adds to the complexity of the language, we can give a rough estimate of the *lexical specialization* of a text by summing, for each noun in the text, the length of the shortest hyperonymy path from the word to a root word of the hyperonymy hierarchy (called *unique beginners* in WordNet's terminology). For example, the hierarchy for *Malonylurea* is

> Malonylurea → acid → compound → substance → object → entity

that gives a specialization score of 6 for *Malonylurea*. Measures based on similar principles, but on different relations, can also be defined for adjectives, verbs, and adverbs. This kind of characterization, however, is too simplistic for our purposes: suffice to say that many simple words are found at the leaves of deep hierarchies, e.g.

> salt (table salt) → flavorer → ingredient → food product → food → substance → object → entity

giving a score of 8 (higher than that for Malonylurea!) for the quite common table salt. On the other hand, if we base our score on the frequency of occurrence of words in standard corpora, we end up measuring again an entropy-based measure, that we have already discussed in the previous section.

Clearly, a more sophisticated approach is needed. Given a certain word w, we define the *active subnet* (up to n) of w the set of all the synsets that can be reached (with a path of length at most n) from any synset containing w in the appropriate syntactic category. We denote this set with $\mathcal{A}_n(w)$. Formally,

$$\begin{aligned} \mathcal{A}_0(w) &= \{s \in \mathcal{W} \mid w \in s\} \\ \mathcal{A}_{n+1}(w) &= \mathcal{A}_n(w) \cup \{s \in \mathcal{W} \mid (s' \to s) \in \mathcal{W} \wedge s' \in \mathcal{A}_n(w)\} \end{aligned}$$

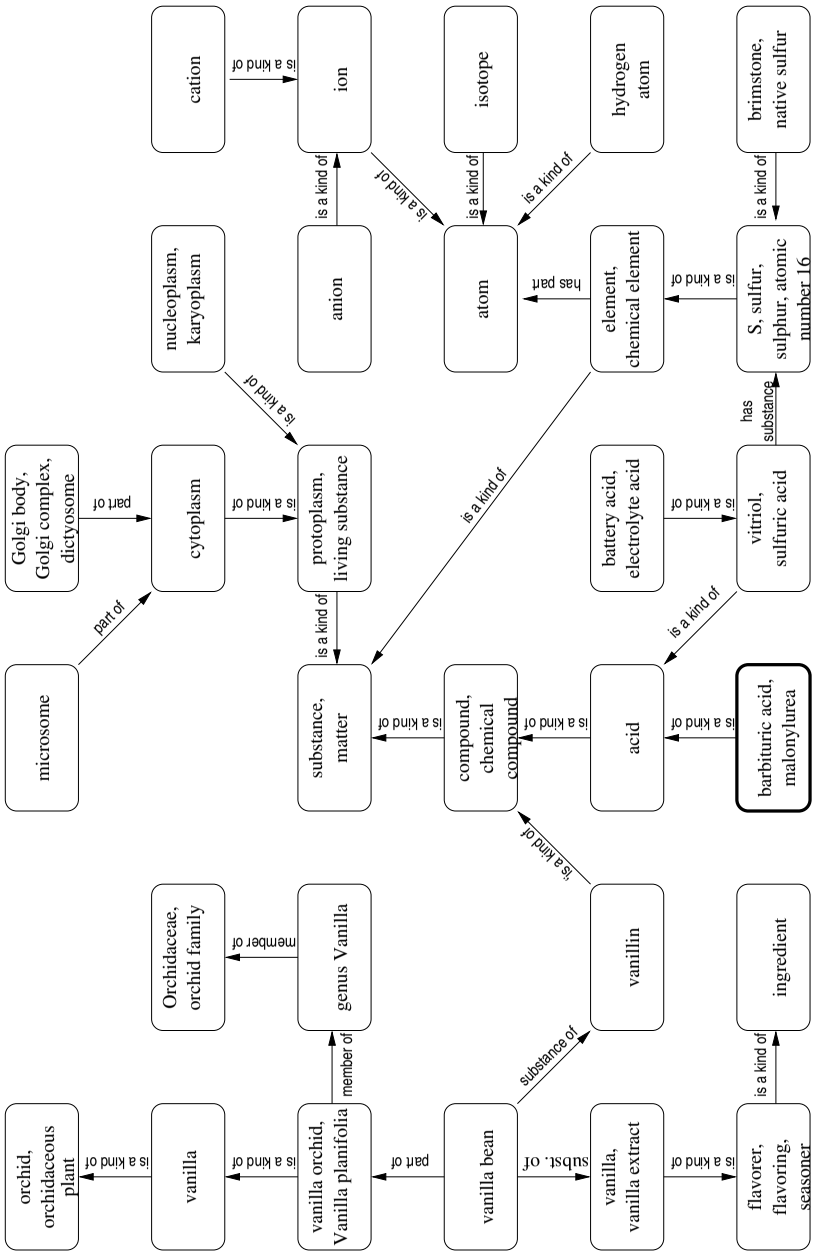


Figure 5: A fragment of the WordNet surrounding *Malonylurea*.

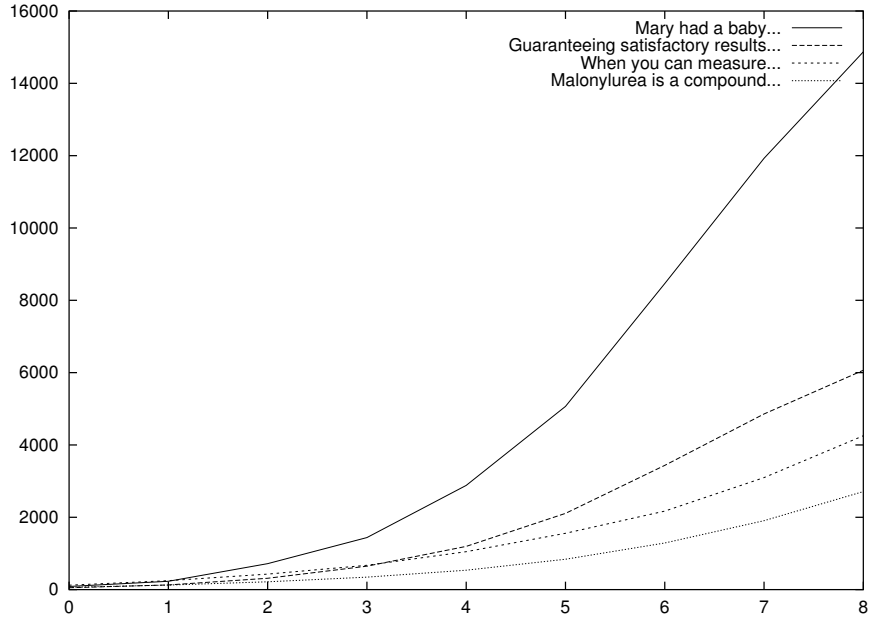


Figure 6: $|\mathcal{A}_n(t)|$ plotted for four sample texts.

where we indicate with $s \in \mathcal{W}$ the fact that synset s appears in the WordNet database, and with $(s' \to s) \in \mathcal{W}$ the fact that the database contains at least a relation linking s' to s. In other terms, $\mathcal{A}_n(w)$ represents the set of all the concepts that are "closely" (depending on n) related to w. We can merge the active subnets of the words in a text t, by computing

$$\mathcal{A}_n(t) = \bigcup_{w \in t} \mathcal{A}_n(w)$$

$\mathcal{A}_n(t)$ represents the set of all the concepts "closely" related to t; due to the presence of the set-union operator, $|\mathcal{A}_n(t)|$ tends to grow rapidly with n when t contains terms from weakly correlated domains, whilst the growth is slower for texts focusing on a well defined domain. So, if t is a scholarly work on the role of barbiturate drugs in the cure of sleep disorders, we can expect $\mathcal{A}_n(t)$ to converge rapidly on the domain of chemistry and medicine, and stabilize there. On the other hand, if t contains magniloquently flourished prose, allegories, or euphemisms, $\mathcal{A}_n(t)$ would probably grow more slowly, but converge to a much larger subset of the whole WordNet.

Figure 6 plots $|\mathcal{A}_n(t)|$, for growing values of n, in the case of four of our sample texts. The measures were obtained by counting the number of synsets in WordNet connected to at least a word of t by a path of length at most n that does not include any hyponymy link.

As can be observed in the figure, what we would judge as simpler texts (like our "Mary had a baby" sample) tend to produce large connected subnets, and thus have higher values for $|\mathcal{A}_n(t)|$, while more complex texts (like our "Malonylurea is a compound" sample) tend to remain confined to smaller, more specialized subnets, and thus have lower values for $|\mathcal{A}_n(t)|$. The two other samples, which intuitively have intermediate complexity, lie between these extremes. This technique appears to capture a significant aspect of *complexity*, that is not easily characterized through other metrics — a facet of complexity that is directly related to the semantic content of a text.

In order to more easily manage this measure, we define a punctual version as follows:

$$\mathcal{PA}(t) = \begin{cases} \frac{3 \cdot 10^3}{|\mathcal{A}_5(t)| - |\mathcal{A}_2(t)|} & \text{if } |\mathcal{A}_2(t)| \neq |\mathcal{A}_5(t)| \\ \\ 0 & \text{otherwise} \end{cases}$$

In other words, we take the reciprocal of the slope of a linear approximation of a fragment of $|\mathcal{A}_n(t)|$ as an approximated indicator for the whole function. The fragment chosen is the part of the curve that lies between the values 2 and 5. This choice, which is not essential, is motivated by the consideration that the semantic link between two words connected by a path in WordNet becomes weaker when the paths become longer. The factor 1000 that appears in the formula is only used to adjust the magnitude of the result. The numeric values of $\mathcal{PA}(t)$ for our sample texts are as follows:

Mary had a baby...	1.131
Guaranteeing satisfactory results...	2.817
When you can measure...	3.746
Malonylurea is a compound...	7.282

Once again, $\mathcal{PA}(t)$ can be taken as it is, as a factor for the absolute *complexity* of a text, or normalized by some measure of *size* (in our case, $|\mathcal{A}_0(t)|$ is a reasonable choice) to measure density-like complexity.

Let us now consider which of our axioms are satisfied by these measures. Axiom Zero is clearly satisfied with $\mathcal{A}_n(\emptyset) = \emptyset$, $|\mathcal{A}_n(\emptyset)| = \mathcal{PA}(\emptyset) = 0$. Axiom 1 is also satisfied. In fact,

$$\forall t, t',\ t \subseteq t' \Rightarrow \mathcal{A}_n(t') = \mathcal{A}_n(t) \cup \mathcal{A}_n(t' \setminus t) \supseteq \mathcal{A}_n(t)$$

it follows that

$$\forall t, t',\ t \subseteq t' \Rightarrow |\mathcal{A}_n(t)| \leq |\mathcal{A}_n(t')|$$

thus proving punctual monotonicity. $\mathcal{PA}(t)$, instead, is not monotonic, since the slopes of two linear functions do not depend on their relative magnitude. As a counter-example, consider the case in which $t \subseteq t'$ with $|\mathcal{A}_2(t)| = 1000$, $|\mathcal{A}_5(t)| = 2000$, and $|\mathcal{A}_2(t')| = 3000$, $|\mathcal{A}_5(t')| = 5000$. Then, $\mathcal{PA}(t) = 3$, but $\mathcal{PA}(t') = 1.5$, thus disproving monotonicity.

Similar arguments hold with respect to compositionality. $\mathcal{A}_n(t)$ is compositional, taking the append operator as T and set union as M:

$$\mathcal{A}_n(T(t,t')) = \mathcal{A}_n(t \cdot t') = \mathcal{A}_n(t) \cup \mathcal{A}_n(t') = M(\mathcal{A}_n(t), \mathcal{A}_n(t'))$$

Unfortunately, neither $|\mathcal{A}_n(t)|$ nor $\mathcal{PA}(t)$ are compositional, due to the fact that the result of the set union operation on subnets depends on the actual overlap between the subnets. To prove that the measures are not compositional, it is sufficient to consider the case $t =$ malonylurea, $t' =$ alkapton: we have $\mathcal{A}_2(t) = \{\text{malonylurea, acid, compound}\}$ and $\mathcal{A}_2(t') = \{\text{alkapton, acid, compound}\}$, thus

$$\begin{aligned}
4 &= |\mathcal{A}_2(t \cdot t')| \\
&= |\mathcal{A}_2(T(t,t'))| \\
&= M(|\mathcal{A}_2(t)|, |\mathcal{A}_2(t')|) \\
&= M(3,3) \\
&= M(|\mathcal{A}_2(t)|, |\mathcal{A}_2(t)|) \\
&= |\mathcal{A}_2(T(t,t))| \\
&= |\mathcal{A}_2(t \cdot t)| \\
&= 3
\end{aligned}$$

The absurd conclusion $4 = 3$ proves that no such M can exist, and thus that $|\mathcal{A}_n(t)|$ (and $\mathcal{PA}(t)$, as a consequence) is not compositional. The first and last equality in the proof above also gives evidence that the measures satisfy classifiability. The definition $\mathcal{A}_n(t) = \bigcup_{w \in t} \mathcal{A}_n(w)$ shows trivially that our measures are unstructural and order-independent, since the active subset of a text t does not depend on the order in which the words w are taken.

Clearly, the very idea of a metric for *complexity* based on the semantic characteristics of a text opens the way to a huge number of possible metrics for special cases. We do not have the space here to even attempt an analysis and a categorization of the possible metrics, but it is worthwhile to remark that semantic metrics can also

Metric	Axiom 0	1	2	3	4	5	6	Scale
Readability								
R, G, G'				X		X	X	interval
Information content								
H on letters & words	X			X		X	X	ratio
H on parts-of-speech	X			X	X		X	ratio
H_n, $n > 0$	X			X	X			ratio
Structural								
$\sum_{s \in t} \mathcal{C}(s)$, where $\mathcal{C}(s) = \lvert N \rvert$, $\max_{n \in L} d_n$, or $\sum_{n \in L} \log_2 d_n$	X	X	X	X	X		X	absolute ($\lvert N \rvert$), ratio (others)
Semantic								
$\lvert \mathcal{A}_n(t) \rvert$	X	X		X		X	X	absolute
$\mathcal{PA}(t)$	X			X		X	X	ratio

Table 5: Axioms satisfied and measurement scales for each of the proposed metrics.

cover structural aspects as well as unstructural ones. For example, [ZGM01] presents a technique for synthesizing a set of logic formulas that is equivalent to a set of natural language *requirements* for a software system. Such a description in logic terms — that is clearly independent from the particular syntax and lexicon used in the original text — can be measured in terms of nodes in a SP-tree representation of the formulas, or in terms of number of conjuncts, etc.

These and similar techniques, however, are too specialized to be applicable to totally general, unrestricted texts on which no assumptions can be made. Still, in many practical applications these *ad hoc* metrics are usually more convenient and more significant than the generic metrics that we have discussed so far.

3.5. *Summary*

It is now time to summarize the various metrics that we have proposed, together with the axioms they satisfy.

As can be observed in Table 5, the various metrics satisfy different axioms, and produce results on different measurement scales. In defining a Factors-Criteria-Metrics system for *complexity*, specific metrics can be chosen among those in Table 5 in such a way as to preserve (if possible) desired properties.

Ideally, a global evaluation of *complexity* should take into account all the facets that we have discussed, and possibly others, by computing an index based on at least a

metric from each group, and by assigning weights to the different measures according to the needs that such an index is intended to satisfy.

In particular, any good characterization of *complexity* should include at least a metric to cover structural aspects (Axiom 4), and at least another one to cover unstructural aspects (Axiom 5). Also, there is a need for some order-dependent factor to intervene in the characterization, to avoid unpleasant results like those we found for *readability* at the end of Section 3.1. As we have seen, most metrics proposed in the literature are order-independent, and satisfy Axiom 6. Sequence-based information content metrics, structural metrics (on order-dependant structures, e.g. syntactic structure), and semantic metrics for special cases are among the few metrics that do take ordering into account; at least one of these should be included in a factor for *complexity*.

While all of the metrics that we have proposed can be easily and efficiently computed by automatic tools, only a few of them have good computational properties, and satisfy axioms 0 (Zero), 1 (Monotonicity), and 2 (Compositionality). Indeed, only structural metrics satisfy all three axioms, and even then, only in few cases, and only for composition functions that are accurately construed according to the specific structures that are being measured.

A good characterization of *complexity* should also include measures on different levels. The entropy measure H on characters can provide an indication of *lexical complexity*, as can the various *readability* metrics. On a higher level, $|\mathcal{A}_n(t)|$ and $\mathcal{PA}(t)$ ignore the particular words used to express a lexical concept, and measure characteristics of the concept itself instead. Entropy measures on parts-of-speech and structural measures on syntactic structure ignore the lexical concepts entirely, and concentrate instead on how statements are built from their constituents. Measures on the argumentative structure disregard English syntax altogether, and only consider the way in which statements are used in building an argument.

Clearly, each of these levels deserves its own notion of *complexity*, and all of them contribute to the global *complexity* of a text.

4. *Conclusions*

At the beginning of this work, we asked ourselves *why* "Mary had a baby" was intuitively simpler than Lord Kelvin's thoughts on measurement. After our discussion, we still do not have a good, single answer to that question.

What we have presented, though, is a set of conceptual, methodological, and technical tools for building answers to that question. The representational theory of measurement provided us with a theoretical framework to understand the essence of measurement and the significance of the results that can be obtained, according to the various measurement scales. The Factors-Criteria-Metrics model suggested how we

could compose disparate facets of *complexity*, and several possible metrics for each facet, into a uniform view. We also presented a number of specific metrics, either already known in the literature or novel, that attempt to measure some of the facets in an objective, repeatable, and automatic way. There is no need to stress here that all these three characteristics are indispensable to obtain metrics that are useful in practice and provide significant measures.

We also developed a system of axioms, that can guide us in selecting the metrics that are more appropriate in certain contexts, while having a precise understanding of their mathematical behavior.

The final step along this path — defining a metric for *complexity* as such — is not taken in this work. The reason is clear: we do not believe that there is a single *complexity* attribute that can characterize a text. Rather, many *complexities* can be defined, (e.g., *lexical complexity*, *syntactic structure complexity*, *argument complexity*, etc.), aimed at measuring different qualities of a text. This view is supported by the observation that all these attributes are largely independent of each other: the same text can feature a varied and flourished lexicon, and at the same time present an elementary syntactic structure. Politician's speeches are commonly acknowledged to present challenging syntactic structures, with very little information content. Four years old children's attempts to convince their parents that it is not bed time YET, can be syntax-impaired and use elementary lexicon, and still make perfectly good and quite complex arguments.

What is left then of our purpose of measuring *complexity*? We can still define a somewhat arbitrary "complexity score", say through a weighted sum of the measures for various facets, much in the same way as a decathlon athlete builds his or her score by running, jumping, swimming, etc. Naturally, the cumulative score does not say much on how good the athlete is at any particular activity, but it is still used conventionally to compare athletes. On the other hand, we could instead accept that *complexity* is not a single number, but a series of numbers. This is what we do with *color*; to specify a color, three numbers are generally used (e.g., red, green, and blue components). These *measurement vectors* are certainly more informative, but more difficult to manage, than arbitrary scores.

We have to leave the final choice to the particular application. Still, a rigorous and mathematically sound method like the one we presented will allow an informed and well thought out decision to be made, whereas in the past measurement of complexity has been often seen, we could say, as The Art of Assigning Random Numbers to Random Phenomena, and Making a Sum in the End. Complexity measurement is a fascinating, thought-provoking and practically relevant subject, and it deserves the most precise and rigorous treatment that we can attain.

The linguistic resources and the software programs developed in support of the metrics proposed here are freely available for academic use. The relevant archives, together with links to other related software packages, can be found on the World Wide Web at the address `http://www.di.unipi.it/~gervasi/TCBook/`.

References

[Abb84] Edwin A. Abbot. *Flatland – A Romance of Many Dimensions*. 1884.

[All95] James Allen. *Natural Language Processing*. Benjamin Cummings, 1995.

[Col65] Edmund B. Coleman. On understanding prose: Some determiners of its complexity. Technical Report GB-2604, National Science Foundation, 1965.

[DC48] Edgar Dale and Jeanne S. Chall. A formula for predicting readability. *Educational Research Bullettin*, 27:11–20, 37–54, February 1948.

[Fel98] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA., 1998.

[FJP51] James N. Farr, James J. Jenkins, and Donald G. Paterson. Simplifications of Flesch reading ease formula. *Journal of Applied Psychology*, 35(5):333–337, October 1951.

[Fle48] Rudolf F. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233, June 1948.

[FP97] Norman E. Fenton and Shari Lawrence Pfleeger. *Software Metrics – A Rigorous & Practical Approach*. International Thomson Computer Press, second edition, 1997.

[Ger01] Vincenzo Gervasi. The CICO domain-based parser. Technical Report 01-25, Dipartimento di Informatica, Università di Pisa, 2001.

[Gun52] Robert Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.

[Har28] R. V. L. Hartley. Transmission of information. *Bell System Technical Journal*, page 535, July 1928.

[Hol68] John B. Holquist. *A Determination of Whether the Dale-Chall Readability Formula may be Revised to Evaluate More Validly the Readability of High School Science Materials*. PhD thesis, Colorado State University, 1968.

[McL69] G. Harry McLaughlin. SMOG grading – a new readability formula. *Journal of Reading*, 12:639–646, May 1969.

[MP82] Douglas R. McCallum and James L. Peterson. Computer-based readability indexes. In *Proceedings of the ACM'82 Conference*, pages 44–48, October 1982.

[MRW77] J. A. McCall, P. K. Richards, and G. F. Walters. Factors in software quality. Technical Report RADC TR-77-369, U.S. Rome Air Development Center, 1977. (volumes I–III).

[MSM93] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19:313–330, 1993.

[Nyq24] H. Nyquist. Certain factors affecting telegraph speed. *Bell System Technical Journal*, page 324, April 1924.

[Nyq28] H. Nyquist. Certain topics in telegraph transmission theory. *A.I.E.E. Transactions*, 47:617, April 1928.

[PSK58] R. D. Powers, W. A. Sumner, and B. E. Kearl. A recalculation of four readability formulas. *Journal of Educational Psychology*, 49:99–105, April 1958.

[Sch94] Helmut Schmid. Probabilistic part–of–speech tagging using decision trees. In *Proceedings of the Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK, September 1994.

[Sha48] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.

[ZGM01] Didar Zowghi, Vincenzo Gervasi, and Andrew McRae. Using default reasoning to discover inconsistencies in natural language requirements. In *Proceedings of the 2001 Australia-Pacific Software Engineering Conference*, Macau, November 2001.