John Grover: "A model for a non-discrete grammar". Master thesis, University of Oslo, february 1992. Excerpt pages 20-27

Basic concepts in probability and information theory.

The probability of an item, such as a letter or a word in printed English, expresses the chance for this item to occur in a sequence of symbols. It is expressed as a number between zero and one, and is the same as the frequency of the item (expressed as a percentage) divided with 100. If 50% of all items (in a sample) is a certain symbol, then its probability - the chance for encountering this symbol - is 0.5. For example, in printed English, if a letter has a frequency of 1 %, then its probability is 0.01.

More precisely, this is the unconditional probability. When we measure it, we take no notice of its surroundings or any other factors which could motivate its occurrence. Thus, the unconditional probability of this letter in printed English is 0.01. If, however, we assume that this is the letter 'u', and go through a text corpus and stop at each letter 'q', and then investigate the frequencies of the letters which follow 'q', we will find that the letter 'u' is heavily overrepresented. In fact, in this specific position, after the letter 'q', the chances for meeting the letter 'u' are very high, probably very close to 1.0. This means that if we take the surrounding into consideration, the probability may be very different from the overall unconditional probability. This probability is termed the conditional probability of 'u' given 'q' (here in the immediately preceding position). The conditional probability is denoted by p(u|q), and for printed English we will have

p(u|q) ≈ 1.0, while the unconditional probability may be around p(u) = 0.01. We will find that the conditional probability will vary with the surroundings and the conditions we specify. For example, we can compute from a corpus the probability that the word 'man' occurs after some specified sequence of words, such as e.g. 'the old'. We will find that p(man|the old) is probably higher than the unconditional probability p(man), and it is most certainly higher than p(man|read), i.e., the chances for 'man' to occur immediately after the word 'read'. The conditional probability thus expresses syntactic properties in a sequence of symbols, and a grammar of a string of symbols can be set up on basis of these conditional probabilities. This kind of grammar is often termed a 'finite state grammar' or a 'Markov grammar', and some of its properties has been investigated in e.g. Chomsky (1956) and Chomsky (1957).

Evidently, if p(a|b) = p(a), then the occurrence of 'a' is not affected by the presence of 'b', and we can say that there is no interdependence between the two events 'a' and 'b'. Dependency between events is in probability theory defined by means of products of probabilities. If we, in some corpus, measure the probability of the sequence of two symbols 'ab', and these are independent of each other such that the occurrence of 'a' is not affected by 'b' and vice versa, then we will have the simple rule

$$p(ab) = p(a) * p(b)$$

i.e., the probability of the compound 'ab' equals the probability of 'a' multiplied with the probability of 'b'. If, however, the occurrences of 'a' and 'b' have some interdependence, such that p(b|a) is different from p(b), or p(a|b) (in this case 'b' given immediately after 'a') is different from p(a), then we have by definition that

$$p(ab) = p(a) * p(b|a) = p(a|b) * p(b)$$

Note that the unconditional probability of the 'a' multiplied with the conditional probability of 'b' is the same as the conditional probability of 'a' multiplied with the unconditional probability of 'b'. These laws are the theoretical definitions of the interdependence between events, but they can easily be verified by investigating the distribution in some corpus. If we know the number of 'ab'-combinations in a corpus, and we know the number of 'a' and the number of 'b', then we can easily compute the conditional probabilities from the latter formula. We can then also set up a simple expression for the degree of dependency between 'a' and 'b' by computing the difference between the unconditional and the conditional probabilities. If this difference is zero, then there is no dependency. If it is small, then we can say that the dependency is weak, and if it is large, such as in the case of the letters 'q' and 'u' in printed English, then we can say that they have a large degree of distributional interdependence.

Probabilities are often interpreted as 'the chance for something to happen' given such-and-such preconditions. Psychologically, this can be expressed as an expectancy: If we, as experienced readers of English, read a text one letter after the other and we come to a 'q', then our expectancy for the following letter to be 'u' is reflected in the high conditional probability p(u|q). If we stop at, say, an 'a', our expectancy for the following letter to be a 'u' is much lower, reflected in the low value for p(u|a). Information theory (developed in full form in Shannon and Weaver (1949)) provides an extension of these measurements of psychological expectancies by means of probabilities, although the simple probability values cannot be utilized right away.

What information theory measures, is the amount of information transmitted in a code. It is important to keep in mind that this does not mean the same as the everyday sense of the word. The term information is here used in a technical sense, and denotes (roughly) the capacity for a code to make semantic distinctions. (We will make this more precise in the following). Information theory is a mathematical model for quantifying structure, and does so by presupposing a paradigmatic and a syntagmatic axis for the exposition of data. Consider an alphabet which consists of only two symbols, A and B. We will use this alphabet to transmit a message about the outcome of an experiment. Let us suppose that there are a total of sixteen possible outcomes of the experiment. To make sixteen distinctions with only a two-letter alphabet, we must make a series of code words which consists of four letters in each: AAAA, AAAB, AABA ... ABBB, BBBB. We have appointed the meanings of these combinations with the receiver of the message in advance. To send the report about one result among sixteen possible ones, we must then use four letters. Now, if we have a somewhat larger alphabet, which consists of the four letters A, B, C and D, we will find that the message will also be shorter. With four letters we can make sixteen distinctions by means of only two letters in combination: AA, AB, AC, AD, BA ... CD, DD. Similarly, if we expand the alphabet to consist of sixteen letters A, B ... P, we will find that it is enough to send only one single symbol from this alphabet to report the result of our experiment.

In general, the larger the alphabet is, the shorter can the message be. In linguistic terms: The larger the paradigm, the shorter can the syntagm be. In our example, it makes sense to say that all these three messages, although the first had four symbols, the second had two, and the third had only one, they all contained the same amount of information. This is the content of the term 'information' as it must be understood in the present context. It concerns the number of semantic distinctions which are inherent in a sequence of symbols, and the fundamental concept is in the relationship between paradigm and syntagm. This relationship is logarithmic, as can be seen from our example:

$$2^4 = 4^2 = 16^1 = 16$$

The base expresses the number of symbols in the paradigm, and the exponent is the number of symbols in the syntagm. It is important to grasp this fundamental relationship between syntagms and paradigms, and it recurs in the extended definition of information value.

If we have a paradigm (an alphabet) of m symbols, then we can make m semantic distinctions with a one-letter message. If we will combine two of these letters in a message, we can take each of the m symbols as the first symbol and combine with each of the m symbols as the second of the two, which means that we can have a total of m in the power of two combinations. If we make three-letter combinations, we will find that each of the m in the power of two combinations (of two-letter combinations) can be combined with all m symbols in the third position, which means that there will be a total of m in the power of three combinations. Thus, with three-letter combinations, we can make m in the power of three semantic distinctions. This means that one such three-letter combination from the m-symbol alphabet is inherently opposed to m in the power of three alternative expressions. In general, with syntagms of p symbols from this alphabet, we can make m in the power of p semantic distinctions.

Consider another alphabet, with n symbols. We can find a number q such that m in the power of p is equal to n in the power of q. In this case, a p-letter syntagm from the alphabet of m symbols can have as many distinct forms as a q-letter syntagm from the alphabet of n symbols, and we will, as above, say that these two syntagms contain the same amount of information, since they are both opposed to equally many paradigmatic alternatives.

Now, any number can be written in logarithmic form with any positive number greater than one as base. It is customary to use 2 as base. In the present study, we will use 'log' to denote a logarithm with base 2. We can now write an equation for the two syntagms with equal amount of information as follows:

$$\begin{aligned} m^p &= n^q \\ 2^{(\log_2 m)\,*\,p} &= 2^{(\log_2 n)\,*\,q} \\ p * \log_2 m &= q * \log_2 n \end{aligned}$$

Information theory thus defines the amount of information I in the first syntagm as p log m, and the information in the second syntagms as q log n. This equation - by the concept of 'equal amount of information' - is the basis for the definition of the technical measurement of information:

I(first syntagm) = p log m

From this we can define the entity 'amount of information per

symbol': Since there are p symbols in the syntagm, the amount of information per symbol will be obtained by dividing with p, which gives:

$$I(\text{per symbol}) = \log m$$

In a code, the information value per symbol equals the logarithm of the number of distinct symbols in its paradigm (alphabet). To see the fundamentally logarithmic relationship between syntagms and paradigms, we can rewrite the above expression as follows:

$$p * \log_2 m = q * \log_2 n$$

$$\frac{p}{q} = \frac{\log_2 n}{\log_2 m}$$

Thus the (size of the) paradigms are related to each other such as the inverse of the logarithms of the (size of the) syntagms.

In this definition of information, it has been assumed that all combinations (all permutations) of the symbols have been possible. I.e., when all combinatorial possibilities are utilized, the code will make use of all of its distinctive potential. In this case, all symbols will also be equally frequent (since all of them are used maximally). This means that if there are m symbols in the paradigm, each of these will have a frequency equalling 100/m%. That is, the probability of occurrence for any of these symbols will be 1/m. For a symbol A, we will have

$$p(A) = \frac{1}{m}$$

$$m = \frac{1}{p(A)}$$

We can insert this in the definition of information $I = \log m$, and we get for a symbol A

$$I(A) = \log \frac{1}{p(A)}$$

$$I(A) = -\log p(A)$$

This is the general definition of information, by which the information value of a symbol (a discrete unit) is defined in terms of its probability of occurrence. It is essential that the probability of a symbol signifies size of paradigm. This is a trivial fact as long as all symbols in a paradigm are equi-probable, but it is possibly less obvious when the symbols are not equiprobable, such as is normal for natural language codes.

This may be illustrated as follows: The probability expresses the average distance between occurrences of a symbol. Assume a code with a five-symbol paradigm. If the symbols are equifrequent, a symbol A will occur on the average in every fifth position:

xxAxxxxAxxxxxAxxxAxxxxAxx

This average distance between its occurrences is identical to the number of symbols in the paradigm. Its probability (in this case 1/5 = 0.2) reflects directly the size of the paradigm. Now assume that we have another sequence of symbols where A occurs as follows:

xxAxxxxxxAxxxxxxxAxxxxxAxxxxxxxAxx

From the average distance (seven symbols) between its occurrences, we will see that there is on the average space for six other symbols to be inserted between every occurrence of A. The symbol A signifies, by its frequency, a paradigm of seven symbols. The distinctive potential in A's frequency is such as it will be in a seven-symbol paradigm. It is, from this point of view, irrelevant whether there is in fact six other symbols in the code, or whether the number is higher or lower. If it is lower, the distinctive potential in the other symbols will be lower, but it is still the same for the symbol A, due to its particular probability of occurrence. This means that in a code with non-equiprobable symbols, such as is typical for linguistic codes, the symbols will have different probabilities and thus different information values.

As we said above, a probability value is not only an expression for frequency or chance of occurrence, but it is also a measure on a psychological expectancy. In our example where A occurs with a certain frequency, if we pick out randomly one symbol from this code, we will (as experienced readers of the code) have a certain expectation as to the chances for A to occur. Information value is often seen as a measure on surprise: If A occurs often, we will not be surprised to find it. If it occurs seldom, we will be much more surprised to find it if we choose a symbol randomly. The surprise is here seen as the same as the distinctive potential in the symbol, and is defined to be measurable by the information value $I(A) = -\log p(A)$.

A note on how this function behaves: The logarithm of a number between zero and one is always a negative number. When, in the definition of information value, we have a minus in front of the logarithm, the value becomes positive. This value will be zero if the probability is 1.0, and it will increase with falling probabilities. The surprise value will thus be zero if the probability is 1.0, i.e., if the event always occurs. The value will approximate infinity when the probability approximates zero, which is also well in line with the notion of surprise value: An event which hardly ever happens has a large surprise value. Information values are normally expressed in bits, which is the unit when the logarithm base is 2.

Sometimes the unit 'Hartleys' can be encountered, in which case the logarithm base is 10.

We have so far looked at the unconditional information value. But the occurrence of symbols in linguistic (and other) codes is often highly determined by context. If we return to our example of the letter 'u' following the letter 'q' in printed English, the surprise value of the 'u' in this context is evidently much lower than elsewhere. Also, if we look at the paradigm of symbol occurrences in the position immediately after 'q', we will find that it is heavily restricted compared to other positions, and the letter 'u' is strongly overrepresented in this position. The distinctive potential in this position is very low. The conditional probability of 'u' signifies a very small paradigm, which motivates the very low conditional information value

$$I(u|q) = - \log p(u|q)$$

The distribution is bound by strong sequential constraints. For the letter 'u', the information value is reduced, but for the 'non-u' letters which can be found to occur in this position in printed English (say, the letter 'a' in a name such as 'Qatar'), the conditional information value will be considerably increased compared to other positions.

Whenever there is a skewedness or constraint in distribution, such that some event is over- or underrepresented in the environment of another event, then the conditional information value will deviate from the unconditional information value. Morpheme structure conditions are typical examples of such constraints in distribution, and their strength can be measured by the conditional information value. They state restrictions on the distribution over a limited interval. Typically, at the end of the interval, the distribution will be more free. Another form of distributional structure can be found in the composition of phonemes. These are often defined as a class of more or less similar sounds. For this class, the paradigms of sound qualities which appear at, say, the beginning, the middle and the end of the phoneme interval will be governed by such constraints: After a sudden and short period of stillness, the set of sound qualities which appear (e.g. [p,t,k]-sound bursts) is very small compared to the entire set of sounds. The distributional constraints over phonemes can be recognized in another form as well: Two non-similar sounds can be classified in the same phoneme if they appear in complementary distribution. This is to say that there are strong restrictions on the environments they can appear in.

The distributional freedom is generally small within segments, and larger across segment borders. This pertains not only to phonological or morphological segments, but will be valid for syntactic phrase segments as well. The set of possible

grammatical categories which can follow a determiner is e.g. smaller than the entire set of categories. Typically, we will place a phrase (a segment) border where the distributional freedom increases, and we will tend to conceive a string of syntactic units over which there are distributional constraints as some phrase segment, at some level or other.