

Plurality in Binary: An Investigation in Markedness and Frequency

An under-investigated phenomenon which is found in various language groups is the presence of a “zero morpheme” for one valuation of a binary feature. This dichotomy exists for a given semantic feature such that one feature is overtly valuated morphologically, while the other feature somehow is valuated without such overt morphological representation or, allegedly, a null morpheme. A set of generalized traits exist among such morphemes which group them together into a single category to be considered for analysis. Among these traits are: a “default” or “unmarked” character, a relatively high perceived frequency, and an atomic contribution to the semantics of the basic proposition it is modifying.

In this paper, I consider that these traits facilitate the prominence of constraints to maintain both brevity of the utterance and the exertion of least effort possible in conveying information. As a result, these dominate constraints which may exist to ensure that all allowed values of a semantic feature have an overt morpheme. These arguments reflect the well-known Principle of Least Effort (Zipf, 1949), a generalized principle across several fields of science which asserts that an entity will exert the minimum amount of energy to accomplish a task—in this case, linguistically conveying information.

The assertion that high frequency would lead to zero morphemes, or default valuation, can be further framed in the concept of entropy, which is a measurement of uncertainty over a probability distribution, or more specifically, the number of binary digits needed to represent all possible outcomes of a probability distribution (Shannon, 1948). Low entropy means a low degree of uncertainty, which would allow us to assume that a specific outcome can be deduced with a greater amount of ease, all other variables being the same. Using data from the Brown Corpus of English, I show that a well-known case of the phenomenon in consideration—null singular number in nouns—exists as an extension of this morpheme’s high relative frequency compared to its counterpart, which would lend values to the same binary number feature.

In the corpus, I compare the counts of all singular noun tokens to all plural noun tokens as well as types. For tokens, there exists an extremely clean 75-25 split between the probabilities that a given noun is singular or plural, which would strongly suggest a “marked” or “non-default” characteristic to the overt plural. This distribution results in an entropy measure of .81, which implies less than a single binary digit to represent a binary semantic distinction. This is further reflected for all noun types, with a 72-28 split between singular and plural nouns, and an entropy of .85.

Because of the binary output of an entropy calculation over the outcomes of a probability distribution, this facilitates the representation of the singular-plural distinction in English as the digits 1 for the overt plural and 0 for the null singular. The advantage to this analogy is the default character of zero—a representation of no value. Because a zero at the $n+1 \dots \infty$ th position beyond the highest-power of the radix of an integer which has a non-null value can be omitted in a numeric representation, it can be argued that it is not an act of deduction that there is no value beyond this highest non-null value, but rather, more probably, that zero is considered a default to all higher powers which are not numerically represented via the act of logical induction.

Such findings would support work in the field of combinatory categorial grammar as seen in Hoeksema (1985). Because of the high frequency of zero morphemes, there is

arguably less processing involved in the computation of an overt semantic assignment as rewriting a default feature than given certain algorithms which would seek to add the missing semantic feature, as seen in Aone and Wittenburg (1990). This data suggests that the phenomenon of “zero morphemes” should in fact be viewed as overt non-default valuation of a binary feature. One solution to this problem is that non-valuated number features for a given noun could easily be valuated at the final step of parsing using finite state methods, so as not to risk overwriting overtly mentioned number features.

Semantically, it can be assumed that a mention offers the presupposition that there exists in some possible world at least one instance to which the mention refers. Let us visualize this assumption as a single non-null binary digit, one to the 2^0 place, or 1. By doing so, we can then visualize that the feature for singular—not more than one—could be represented as a 01, and therefore, as described above, roughly equivalent to the mention itself, 1. Both the plural morpheme *-s* and any digit greater than one would increase these respective representations greater than their zero counterparts. In this case, even simply more than one, rather than at least one, would result in the binary representation of number for a proposition to be 10. Because of this relation, plurality is therefore a monotone feature, asserting that something is true for the set of all things for which there exist greater than one instance, which is a proper subset of all things which there exist at least one instance.

Building off of this perspective that plurality is the specification of the mentioned members of a set as a proper subset of those which are by mentioned default, and given the statistical and numerical arguments proposed above, this work offers a new perspective to the interface between morphology, semantics, and how generalization over a set of data can lead to default assumptions in meaning. Such conceptualizations could be implemented usefully in several fields, ranging from machine modeling of world knowledge and semantic feature generation of raw text to modeling of language acquisition.

References:

C. Aone and K. Wittenburg. Zero morphemes in unification-based combinatory categorial grammar. In *28th Annual Meeting of the Association for Computational Linguistics*, pages 188–193, University of Pittsburgh, Pittsburgh, 1990.

Jack Hoeksema. *Categorial Morphology*. Garland Publishing, Inc., New York and London, 1985.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–432, 623–656, 1948.

George Kingsley Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, 1949.