Authors: James P. Blevins; Farrell Ackerman; Robert Malouf
Affiliations: University of Cambridge; UCSD; San Diego State University

An entropy-based measure of morphological information

1. Introduction

Traditional approaches to morphology tend to treat inflectional systems not as unstructured sets of forms with shared stems or roots but as structured networks of elements. The interdependency of elements is, as Matthews (1991: 197) notes, 'the basis of exemplary paradigms' in the classical grammatical tradition. Although the exemplary patterns and leading forms of traditional descriptions bring out the structure of inflectional systems, traditional accounts are deficient – or at least incomplete – in a number of important respects. In particular, there is no method for measuring the implication structure of a set of forms or, no means of gauging the diagnostic value of specific forms within a set, and no generally accepted way even of identifying the leading forms of a system.

The approach outlined in this talk proceeds from the observation that implicational structure involves a type of **information**, specifically information that forms within a set convey about other forms in that set. Information in this sense corresponds to reduction in **uncertainty**. The more informative a given form is about a set of forms, the less uncertainty there is about the other forms in the set. In inflectionally complex languages, a speaker who has not encountered all of the forms of a given item is faced with some amount of uncertainty in determining the unencountered forms. If the choice of each form were completely independent, the problem of deducing unencountered forms would reduce to the problem of learning the lexicon of an isolating language. However, in nearly all inflectional systems, there are at least some forms of an item that reduce uncertainty about the other forms of the item. Once these notions are construed in terms of uncertainty reduction, the problem of measuring implicational structure and diagnostic value is susceptible to well-established techniques of analysis. The uncertainty associated with the realization of a paradigm cell correlates with its **entropy** (Shannon 1948) and the entropy of a paradigm is the sum of the entropies of its cells. The implicational relation between a paradigm cell and a set of cells is modelled by **conditional entropy**, the amount of uncertainty about the realization of the set that remains once the realization of the cell is known. The diagnostic value of a paradigm cell correlates with the **expected conditional entropy** of the cell, the average uncertainty remains in the other cells once the realization of the cell is known.

2. Information theoretic assumptions

In order to quantify the interrelations between forms in a paradigm, we will use the information theoretic notion **entropy** as the measure of predictability. This permits us to quantify "prediction" as a change in uncertainty, or information entropy (Shannon 1948). The idea behind information entropy is deceptively simple: Suppose we are given a random variable X which can take on one of a set of alternative values $x_1, x_2, \ldots, x_n$ with probability $P(x_1), P(x_2), \ldots, P(x_n)$. Then, the amount of uncertainty in X, or, alternatively, the degree of surprise we experience on learning the true value of X, is given by the entropy $H(X)$:

$$H(X) = -\sum_{x \in X} P(X) \log_2 P(X)$$

The entropy $H(X)$ is the weighted average of the **surprisal** $-\log_2 P(x_i)$ for each possible outcome x_i. The surprisal is a measure of the amount of information expressed by a particular outcome, measured in bits, where 1 bit is the information in a choice between two equally probable outcomes. Outcomes which are less probable (and therefore less predictable) have higher surprisal. Specifically, surprisal is 0 bits for outcomes which always occur ($P(x) = 1$) and approaches ∞ for very unlikely events (as $P(x)$ approaches 0). The more choices there are in a given domain and the more evenly distributed the probability of each particular occurrence, the greater the uncertainty or surprise there is (on average) that a particular choice will be made among competitors and, hence, the greater the entropy. Conversely, choices with only a few possible outcomes or with one or two highly probable outcomes and lots of rare exceptions have a low entropy. One can also quantify the degree of prediction between cells using entropy. The average uncertainty in one variable given the value another is the **conditional entropy** $H(Y|X)$. If $P(y|x)$ is the conditional probability that $Y = y$ given that $X = x$, then the conditional entropy $H(Y|X)$ is:

$$H(Y|X) = -\sum_{x \in X} P(x) \sum_{y \in Y} P(y|x) \log_2 P(y|x)$$

3. Implicational structure in Uralic

To demonstrate how an information-theoretic approach calculates the relative diagnosticity of words, the talk presents morphological patterns of ascending levels of complexity. The inflectional paradigms of Uralic languages are instructive because of the way that they realize inflectional properties by distinctive combinations of stem alternations and affixal exponence. Hence these systems are not amenable to a standard head-thorax-abdomen analysis in which lexical properties are expressed by the root, morphological class properties by stem formatives, and inflectional properties by inflectional affixes.

3.1 Northern Saami

First declension nouns in Northern Saami may inflect according to either of the patterns in Table 1. In nouns of the 'weakening' type, the nominative and illative singular and the essive are all based on the strong stem of a noun, and the remaining forms are based on the weak stem. Nouns of the 'strengthening' variety exhibit a mirror-image pattern, in which the nominative and illative singular and essive are based on the weak stem, and other forms are based on the strong stem. Strong forms, which are set in bold in Table 1, contain a geminate consonant which corresponds to a non-geminate in the corresponding weak forms.

Table 1: Gradation in first declension nouns in Saami (Bartens 1989: 511)

	'Weakening'		'Strengthening'	
	Sing	Plu	Sing	Plu
Nominative	**bihttá**	bihtát	baste	**basttet**
Gen/Acc	bihtá	bihtáid	**bastte**	**basttiid**
Illative	**bihttái**	bihtáide	bastii	**basttiide**
Locative	bihtás	bihtáin	**basttes**	**basttiin**
Comitative	bihtáin	bihtáiguin	**basttiin**	**basttiiguin**
Essive	**bihttán**		basten	
	'piece'		'spoon'	

Given the paradigm in Table 1, we can calculate the conditional entropy of any one cell given any other cell. Take the nominative singular and the locative plural. Each has two possible realizations, and the entropy of each is 1 bit. To find the joint entropy, we look at the four possible combinations of realizations:

Nom Sg	Loc Pl	P
strong	strong	0.0
strong	weak	0.5
weak	strong	0.5
weak	weak	0.0

There are two equally likely outcomes, and the joint entropy is 1 bit. So the conditional entropy, $H(\text{Loc Pl}|\text{Nom Sg})$, is 0 ($H(\text{Nom Sg, Loc Pl}) - H(\text{Nom Sg})$).

That is, knowing the nominative singular realization for a particular lexeme completely determines the realization of the locative plural. We could repeat this calculation for any pair of cells in the paradigm and we would get the same result, as Saami nominal inflection is a completely symmetric system.

3.2 Finnish

The Finnish sub-paradigm in Table 2 illustrates a more typical pattern, in which different **combinations** of cells are diagnostic of declension class membership.

The implicational structure of the paradigms in Table 2 is set out in Table 3. The row expectation $E[\text{row}]$ is the average conditional entropy of a column given a particular row. This is a measure of the **predictiveness** of a form. By this measure, the partitive singular is the most predictive form: if we know the partitive singular realization for a lexeme and want to produce on other paradigm

Table 2: Finnish *i*-stem and *e*-stem nouns (Buchholz 2004)

Nom Sg	Gen Sg	Part Sg	Part Pl	Iness Pl	
ovi	oven	ovea	ovia	ovissa	'door' (8)
kieli	kielen	kieltä	kieliä	kielissä	'language' (32)
vesi	veden	vettä	vesiä	vesissä	'water' (10)
lasi	lasin	lasia	laseja	laseissa	'glass' (4)
nalle	nallen	nallea	nalleja	nalleissa	'teddy' (9)
kirje	kirjeen	kirjettä	kirjeitä	kirjeissä	'letter' (78)

Table 3: Conditional entropy $H(\text{col}|\text{row})$ of Finnish *i*-stem and *e*-stem nouns

	Nom Sg	Gen Sg	Part Sg	Part Pl	Ines Pl	$E[\text{row}]$
Nom Sg	—	1.333	1.667	0.874	0.541	1.104
Gen Sg	0.459	—	0.459	0.459	0.459	0.459
Part Sg	0.333	0.000	—	0.333	0.333	0.250
Part Pl	0.333	0.792	1.126	—	0.000	0.563
Ines Pl	0.459	1.252	1.585	0.459	—	0.939
$E[\text{col}]$	0.396	0.844	1.209	0.531	0.333	0.663

cell chosen at random, we will require only 0.250 bits of additional information on average. In contrast, given the nominative singular, we would need an addition 1.104 bits of information on average. The column expectation $E[\text{col}]$ is the average uncertainty given a row remaining in a particular column. In contrast to the row expectations, this is a measure of the **predictedness** of a form. By this measure, the inessive plural is the most predicted form: if we want to produce the inessive plural for a lexeme and know some randomly selected other form, we will require on average another 0.333 bits of information.

Analyses of noun declensions in Tundra Nenets further confirm the value of entropy measures as a gauge of implicational structure and diagnostic value.

References

Bartens, H.-H. (1989). *Lehrbuch der saamischen (lappischen) Sprache*. Helmut Buske Verlag.

Buchholz, E. (2004). *Grammatik der finnischen Sprache*. Bremen: Hempen Verlag.

Matthews, P. H. (1991). *Morphology*. Cambridge University Press.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379–423, 623–656.