Discreteness and gradience as co-operating techniques in modelling a language

Jan Křivan
Charles University in Prague

The paper deals with linguistic modelling. From a methodological viewpoint, it concerns the role of categorization in a metalinguistic domain of linguistic description and explanation. More specifically, the relation between discreteness and gradience in language modelling is surveyed.

In general, I claim that every sort of categorization – observing clearly defined rules – has a potential to depict relevant language properties. There is no unique, principally correct or false way of categorization. The choice of categories and the ways of their construction depends chiefly on stipulated scientific aims. The role of every linguist is to seek and describe relevant categories. However, it is not obvious, what a relevant category looks like. Some scholars believe in universal categories, whereas others take only language-particular categories for granted. Linguists have very different responses to the basic question *What should a model of (a) language tell us about language?*

In this paper, I pursuit a descriptive model of language which is functional in nature and concerns a particular language. It doesn't presuppose any pre-established categories (in the sense of Haspelmath 2007). It should be able to capture language-particular facts in a comprehensible way and make them available to functional explanations, e.g. in terms of frequency, as in functional usage-based approaches (Bybee 1995, 2007, Haspelmath 2008a). A model of a particular language should embrace and generalize (i.e. describe) data of two different kinds:

(i) "raw" linguistic data (e.g. corpus data, fieldwork data); mainly by means of distributional analysis, i.e. generalizations over texts, this part of a model corresponds roughly to Basic Linguistic Theory (e.g. Dixon 2010);

(ii) "metalinguistically processed" linguistic data; besides linguists' introspection which is inevitable for (i), it should capture language facts in corespondence to the results of grammatical judgments and other psychological experiments, i.e. generalizations over psychologies.

A model of a language is viewed as generalized empirical knowledge of language use and language processing. In other words it should describe adequately "what we produce" by the means of "what we think that we produce".

Accordingly, the data should be linguistically generalized and categorized both as gradient and discrete. The reason consists in (ii): generally, we perceive the language and its categories psychologically in both ways: a burden of empirical evidence in experimental psychology since Rosch's first investigations has shown that a gradient way of perception and modelling is a commonly used strategy by humans. In linguistics, Aaarts (2004) showed that gradience is compatible with discrete modelling. For these reasons, both techniques are meaningful. Some grammatical categories can be captured better using gradience, other ones using only discreteness. Gradience functions in a descriptive model as a linguistic instrument.

The functional nature of a model consists in its explanatory potential. It reckons with different corroborating variables external to the model, e.g. frequency, processing ease, prediction, economy.

In addition, a descriptive model is open to be accompanied by hypothesized cognitive processes. These processes don't serve as an explanation itself, but they can indicate the way in which some of the explanatory variables might be linked to a language model via cognition (i.e. cognitive processes are mediating these explanations).

Empirically, I will illustrate the issue by means of selected data gathered in my projects focused on (i) the description of possession in Czech and (ii) on the frequency analysis of comparative and superlative forms in the Czech National Corpus.

Example (1):

(a) A descriptive model: In Czech, there are in principle three ways how to form an adjective comparative (non-reduced suffixation: *ostr-ý – ostř-ejší* 'sharp – sharper' (-ejší), reduced suffixation: *krátk-ý – krat-ší* 'short – shorter' (-ší), suppletion: *dobr-ý – lep-ší* 'good – better') – corresponding to the three stages (inflection – derivation – lexical) of a synchronic scale of fusion (Bybee, Perkins, Pagliuca 1994:41).

(b) An explanation: Absolute high frequency of use (e.g. Bybee 2007) explains why linguistic forms tend to be reduced. Reduced comparative forms tend to be highly frequent. Suppletive forms have the highest frequency.

(c) Cognitive processes show possible cognitive links between reduced forms in a model and frequency. E.g. Haspelmath's (2008b) explanation of reduced forms due to predictability for the hearer or Bybee's (e.g. 2007) explanation due to speaker's routinization.

As a result, additional external properties can be optionally assigned to the comparative forms in a descriptive model: more frequent comparative forms tend to be lexicalized, less frequent forms tend to be regular.

The conclusion of the paper is: both gradient and discrete categories are suitable to play, side by side, a significant role in a descriptive model of a language and can be used for external functional explanations.

References:

Aarts, Boas. 2004. Modelling linguistic gradience. *Studies in Language* 28:1, 1–49.
Bybee, Joan. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10, 425–455.
Bybee, Joan. 2007. Introduction. In: Bybee, Joan. *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press, 5–22.
Bybee, Joan, Revere Perkins & William Pagliuca. (1994) *The evolution of grammar: tense, aspect and modality in the language of the world*. Chicago: University of Chicago Press.
Dixon, R. M. W. 2010. *Basic linguistic theory: Volume 1, Methodology*. Oxford: Oxford University Press.
Haspelmath, Martin. 2007. Pre-established categories don't exist: consequences for language description and typology. *Linguistic Typology* 11.1, 119–132.
Haspelmath, Martin. 2008a. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19, vol. 1, 1–33.
Haspelmath, Martin. 2008b. Reply to Croft and Haiman. *Cognitive Linguistics* 19, vol. 1, 59–66.