Balthasar Bickel
Universität Leipzig
Institut für Linguistik
Beethovenstrasse 15
04107 Leipzig, Germany
+49-341-973 7609
bickel@uni-leipzig.de

ABSOLUTE AND STATISTICAL UNIVERSALS

Language universals are statements that are true of all languages, for example: "all languages have stop consonants". But beneath this simple definition lurks deep ambiguity, and this triggers misunderstanding in both interdisciplinary discourse and within linguistics itself. A core dimension of the ambiguity is captured by the opposition "absolute vs. statistical universal", although the literature uses these terms in varied ways. Many textbooks draw the boundary between absolute and statistical according to whether a sample of languages contains exceptions to a universal. But the notion of an exception-free sample is not very revealing even if the sample contained all known languages: there is always a chance that an as yet undescribed language, or an unknown language from the past or future, will provide an exception.

It is impossible, in principle, to survey all languages of our species. If we nevertheless want to make claims about all languages, only two routes are open: *a priori* deduction of necessarily true statements, or statistical extrapolation from empirical samples to the entire set. Absolute universals can then be defined as those that are necessarily true, statistical universals as those that are extrapolated from samples.

Absolute Universals

For statements to be necessarily true, they must follow from *a priori* assumptions. The assumptions that linguists make are diverse and heavily debated. An example is the assumption that WORDS consist of MORPHEMES, i.e. minimal form-meaning pairs. If one accepts this, then it is necessarily true that all languages have morphemes, and there cannot be exceptions. Why? Suppose someone claims to have discovered a language without morphemes. One can of course simply analyze the language without mentioning morphemes, but obviously that cannot challenge the universal just because one can always defend it by re-analyzing the language *with* morphemes. The only true challenge would be to show that analyzing some data in terms of morphemes leads to structures that are in conflict with other assumptions, e.g. that form-meaning pairs combine exclusively by linear concatenation. The conflict can be illustrated by languages with morphologies like the English plural *geese*, where the meanings "plural" and "goose" do not correspond to linear strings of morphemes. Confronted with such data, there are three options: (i) give up the notion of morpheme; (ii) give up the assumption of linear concatenation; (iii) add additional assumptions that reconcile the conflict. On any of these options, the universal remains exceptionless: on solution (i), no

language has morphemes; on solutions (ii) and (iii), all languages have morphemes. As a result, absolute universals can never be falsified by individual data. Their validity can only be evaluated by exploring whether they are consistent with other absolute universals that are claimed simultaneously.

Absolute universals can also be thought of as those aspects of one's descriptive metalanguage — often called a "theoretical framework" — that are necessarily referred to in the analysis of every language, i.e that constitute the descriptive *a priori*. Depending on one's *a priori*, this includes, apart from the morpheme, such notions as distinctive feature, constituent, argument, predicate, reference, agent, speaker, etc. In some metalanguages, the *a priori* also includes more specific assumptions, e.g. that constituents can only be described by uniform branching (all to the left, or all to the right), or only by binary branching, etc.

The status of absolute universals is controversial. For many linguists, especially in TYPOLOGY and HISTORICAL LINGUISTICS, absolute universals are simply the descriptive *a priori*, with no additional claim on biological or psychological reality. The choice between equally consistent universals/metalanguages — e.g. between options (i) – (iii) in the example above — is guided by their success in describing structures and in defining variables that capture distributional patterns — an evaluation procedure comparable to how technical instruments for analyzing objects are evaluated in the natural sciences. In the morphology problem from before, typologists would most likely chose option (ii) — because it allows defining a variable of stem-internal vs. affixal plural realization that has an interesting distribution (suggesting, for example, that within-stem realization is favored by a few families in Africa and the Near East).

In GENERATIVE GRAMMAR, by contrast, absolute universals are not only thought of as descriptively *a priori* but also as biologically given in what is called UNIVERSAL GRAMMAR: they are claimed to be innate and to be identical to the generalizations that a child makes when learning language. Thus, if the morpheme is accepted as a universal, i.e. *a priori* term of our metalanguage, it will also be claimed to be part of what makes languages LEARNABLE and to be part of our genetic endowment. An immediate consequence of such an approach is that something can be claimed as universal even if it is not in fact necessary in the analysis of every language. For example, even if some language (e.g. the Rotokas language of Bougainville) lacks evidence for nasal sounds, one could still include a distinctive feature [±nasal] in Universal Grammar. Rotokas speakers are then said to have the feature as part of their genetic endowment even if they don't use it.

This view of absolute universals is highly controversial: many linguists limit absolute universals to what is descriptively necessary in every language; many psychologists propose that children apply different and much more general principles in AQUIRING A LANGUAGE than those found in linguists' metalanguages; and to date no absolute universal has been confirmed by genetic research.

Statistical Universals

What is not an absolute universal is a variable (or character, or PARAMETER): some languages have a certain structure or they don't have it, or to different

degrees. Interestingly, most variables show some skewing in their distribution: some values of a variable are favored only in certain geographical areas (relative pronouns in Europe) or only in certain families (stem-internal inflection in Afroasiatic). But some values are globally favored (e.g. nasals), or, what is more typical, globally favored under certain structural conditions (e.g. postnominal relative clauses among languages with objects following the verb). These global preferences are called unconditional (unrestricted) and conditional (restricted) statistical universals, respectively. (An alternative term for conditional statistical universals is *IMPLICATIONAL UNIVERSAL*, but this invites confusion because their probablistic nature differentiates them from logical implications; cf. Cysouw 2005)

Statistical universals are mostly motivated by theories of how languages develop, how they are used, how they are learned, and how they are processed. One such theory, for example, proposes that processing preferences in the brain lead to a universal increase in the odds for postnominal structures among verb-object languages (Hawkins 2004).

Statistical universals take the same forms as statistical hypotheses in any other science — for example, they can be formulated in terms of regression models. They can be tested with the same range of statistical methods as in any other science, and, again as in other sciences, the appropriate choice of models, population assumptions, and testing methods is an issue of ongoing research (e.g. Cysouw 2005, Janssen et al. 2006, Maslova 2007).

A central concern when testing statistical universals is to ascertain true globality, i.e. independence of area and family. Areas can be controlled for by standard factorial analysis, but it is an unsettled question just what the relevant areal relations are: for example, should one control for the influence of Europe or the entire Eurasia or both? A quick solution is to assume a standard set of five or six macro-areas in the world and accept as universal if a distribution is independent of these (Dryer 1989). But the rationale for such a set is problematic, and this has lead to a steep surge of interest in research on areas and their historical background (e.g. Nichols 1992, Haspelmath et al. 2005).

Controlling for family relations poses another problem. Under standard statistical procedures one would draw random samples of equal size within each family and then model families as levels of a factor. However, over a third of all known families are isolates, containing only one member each. And picking one member at random in larger families is impossible if at the same time one wants to control for areas (e.g. admitting an Indo-European language from both Europe and South Asia). In response to this problem, typologists seek to ensure representativity of a sample not by random selection within families but by exhaustive sampling of known families, stratified by area. In order to then control for unequal family sizes, one usually admits only as datapoints per family as there are different values on the variables of interest (Dryer 1989, Bickel 2007).

Samples that are not based on random sampling do not support parametric inference by statistical tests. An alternative to this is randomization methods (Janssen et al. 2006): the null hypothesis in these methods is that an observed preference can be predicted from the totals of the sample (e.g. that an observed 90% postnominal relatives in VO languages could be predicted if 90% of the entire sample had postnominal relatives) — not that the sample stems from a

population without the observed preference. Extrapolation to the total population (the entire of set of human languages) can then only be based on plausibility arguments: if a preference significantly deviates from what is expected from the totals of the observed sample, it is likely that the preference holds in all languages. A key issue in such argumentation is whether the tested variables are sufficiently unstable over time so that a present sample can be assumed to not reflect accidental population skewings from early times in prehistory (Maslova 2000). In response to this, typologists now also seek to test universals by sampling language changes instead of language states — a move that is sometimes called the "dynamization" of typology (Greenberg 1995, Croft 2003).

While the number of proposed statistical universals is impressive — the Universals Archive at Konstanz has collected over 2000 (Plank & Filimonova 2000) —, very few of them have been rigorously tested for independence of area, family, and time.

— Balthasar Bickel

Works cited and suggestions for further reading

Bickel, B. 2007. A refined sampling procedure for genealogical control. *Sprachtypologie und Universalienforschung #, ## - ##.*

Croft, W. 2003. *Typology and universals.* Cambridge: Cambridge University Press [2nd edition].

Cysouw, M. 2005. Quantitative methods in typology. In Altmann, G., R. Köhler & R. Piotrowski (eds.) *Quantitative linguistics: an international handbook, 554 – 578.* Berlin: Mouton de Gruyter.

Cysouw, M. (ed.) 2007. Special issue on analyzing the *World Atlas of Language Structures. Sprachtypologie und Universalienforschung ##.*

Dryer, M. S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13, 257 – 292.

Greenberg, J. H. 1995. The diachronic typological approach to language. In Shibatani, M. & T. Bynon (eds.) *Approaches to language typology,* 143 – 166. Oxford: Clarendon.

Haspelmath, M., M. S. Dryer, D. Gil & B. Comrie (eds.) 2005. *The world atlas of language structures.* Oxford: Oxford University Press.

Hauser, M. D., N. Chomsky & W. T. Fitch 2002. The faculty of language: what it is, who has it, and how did it evolve? *Science* 298, 1569 – 1579. [This paper and the response by Pinker & Jackendoff 2005 launched an ongoing debate on the nature and extent of absolute universals in generative grammar.]

Hawkins, J. A. 2004. *Efficiency and complexity in grammars.* Oxford: Oxford University Press.

Janssen, D., B. Bickel & F. Zúñiga 2006. Randomization tests in language typology. *Linguistic Typology* 10, 419 – 440.

Maslova, E. 2000. A dynamic approach to the verification of distributional universals. *Linguistic Typology* 4, 307 – 333.

Maslova, E. 2007. Meta-typological distributions. *Sprachtypologie und Universalienforschung #, ## – ##.*

Newmeyer, F. J. 2005. *Possible and probable languages: a generative perspective on linguistic typology*. New York: Oxford University Press.

Nichols, J. 1992. *Language diversity in space and time*. Chicago: The University of Chicago Press.

Pinker, S. & R. Jackendoff 2005. The faculty of language: what's special about it? *Cognition* 95, 201 – 236.

Plank, F. & E. Filimonova 2000. The Universals Archive: a brief introduction to prospective users. *Sprachtypologie und Universalienforschung* 53, 109 – 123.

Plank, F. (ed.) 2007. Special issue on the state of typology. *Linguistic Typology* 11.