In: Austin, Peter, Helen Dry and Peter Witternburg [eds.] *Proceedings of the International LREC Workshop on Resources and Tools in Field Linguistics, Las Palmas, 26 - 27 May 2002.* Nijmegen: ISLE and DOBES.

Autotypologizing Databases and their Use in Fieldwork

Balthasar Bickel[#] and Johanna Nichols[*]
[#]University of Leipzig, [*]University of California, Berkeley
Address: Institut für Linguistik, Universität Leipzig, Brühl 34-50, 04109 Leipzig, Germany
E-Mail: bickel@rz.uni-leipzig.de, jbn@socrates.berkeley.edu

Abstract

This paper presents two database methods for crosslinguistic data collection and comparison: autotypologizing and exemplar-based sampling. Autotypologizing dispenses with a priori defined comparative grids and instead lets structural types emerge inductively through a type list that is constantly updated in response to languages entered in a database. Examplar-based sampling allows identification of a single representative of cross-linguistically heterogeneous structural domains such as case. These two methods are helpful tools in fieldwork. Autotypologizing generates inventories of known types. These inventories update researchers' expectance range for newly encountered types (like published typological surveys, but more dynamically). Examplar-based sampling is useful for writing typological profiles at very early stages of description.

The Autotypologizing Method

Traditional typological databases first define a set of crosslinguistic types in a functionally defined domain — in more anthropological terms an *etic grid;* in more psychological terms a *conceptual space* — and then assign each language in the sample to a type. One problem of this procedure is that the initial definitions make it unlikely that hitherto unknown types can ever be discovered: if a type does not fit the original scheme, it is usually treated as a transitional phenomenon, although outside the a priori typology it may as well be a primitive type of its own. An example is what Hale (1976) termed the Adjoined Relative Clause in Australian languages, which in a typology of relative constructions might be analyzed as a borderline or non-prototypical case of relativization although it is a central and well-established construction within the individual grammars. The same problem has been noted in the typology of color terms by Lucy (1997): a common procedure in this domain uses a standardized color chart as an etic grid and then typologizes lexemes on their focal color value. This procedure precludes discovery of lexemes that referentially denote color by means of a semantics based on other concepts than hue, saturation and brightness. A system like the well-known Hanunóo color terms (Conklin, 1955), which conceptualizes 'color' in terms of relative dryness, freshness and lightness, will be either misanalyzed or treated as a non-prototypical instance of color terms. But within Hanunoo there is evidence neither for English-type color values nor for non-prototypicality.

The essential but most difficult challenge in fieldwork is to uncover such local categories and constructions in their own terms. In this task, etic grids are no more helpful than applying Latin-based grammar and meaning concepts to other languages: in the best case, one will get an idea of how the language does not fit the grid; in the worst case, the language is made to fit the grid.

Autotypologizing databases seek to remedy these shortcomings by not assuming any a priori list of cross-linguistic types and by not limiting domains. Instead, the inputter selects from an editable menu of types; the menu provides a complete list of the types so far attested, and when the menu is edited a new type is added and receives a cross-linguistically viable definition. Types can be expanded in any direction that a language takes one. If one starts with relativization types, and a newly entered language has a construction of the adjoined Australian type, that construction will constitute a new type and the functional domain will be expanded and redefined. We call this procedure *autotypologizing* (see http://socrates.berkeley.edu/~autotyp).

Autotypologizing databases were originally designed to improve and expedite cross-linguistic typological work, but they also prove to have descriptive applications. The type lists that emerge from autotypologizing data collection are inventories of expectable types and can give the analyst ideas of what kind of structure may be found in a language. When encountering a construction or category that looks different from what the fieldworker's personal typological knowledge can accommodate, looking up an autotypologized inventory will give him or her immediate access to similar types in other parts of the world (including references, and in many cases even examples with discussion). Or, the list will provide ideas about how far the newly found type deviates from everything else, and how it could or should be defined as a type on its own.

Even at an early stage of data collection these inventories are large enough to cover the great majority of expectable types; they are somewhat more reliable, and available at a much earlier stage of work, than published typological surveys. We are now working on publishing stand-alone copies of the large AUTOTYP database containing the current type lists and a small corpus of cross-linguistic data representing diverse well-described languages. The fieldworker can enter data into his or her copy, or just use the database as an information file without further data entry, and either way systematically explore the typological profile of the language under investigation in a systematic way.

An example of an autotypologizing database surveys NP structures (and is being developed within the AUTOTYP research program). In the first several languages we found three major NP types: construct state, governed cases or adpositions, dependent-driven agreement. After entering more languages, more types emerged and were added to the menu: externally-driven case agreement, incorporation, anti-construct state, etc. Each entry is given a cross-linguistically applicable definition. Examples are:

- *Incorporation*: the dependent is regularly incorporated into the head noun. Example from Kiowa (Watkins & McKenzie, 1984:107)

  (1) nɔ́ː-tɔ̀ː-cègùn (my-brother-dog) 'my brother's dog'

- *Anti-construct state marking* registers the presence of a head on its adjacent dependent but neither agrees with the head nor constitutes case marking. Examples from Tamazight Berber (Penchoen, 1973:19, 54, 62, 64; AC = anti-construct):

  (2) aryaz 'man' (normal state)
  (3) idda uryaz (he.has.gone man.AC) 'the man has gone'
  (4) axam uryaz (tent man.AC) 'the man's tent'
  (5) nnigh uryaz (above man.AC) 'above the man'

  A noun is in the anti-construct when it is a dependent of the preceding word, whether as subject of verb as in (3), as possessor of head noun as in (4), or as object of preposition as in (5). A topicalized and therefore pre-verbal subject is not in the anti-construct state. In other languages, such as Meithei (Chelliah, 1997), the anti-construct marker characterizes different sorts of attributes, including nominalized verbs, numerals, and demonstratives. In each instance, the marker signals the presence of a head noun.

This procedure is time-consuming in the beginning because each new type requires review (and possibly revision) of all previous entries, but after a few dozen languages, new types become less likely to emerge and the typology stabilizes. In the NP structure database this happened after about 40 languages were entered.

Overview of the AUTOTYP Database

The database is a set of 38 fully relational File Maker Pro™ files, or modules, plus bibliography, manuals, map-making tools, and log. The usual form of data entry for all of the essential descriptive points is choosing from a fully relational menu of types so far found. The menu is itself a File Maker Pro file in which each type is a record with an ID number, a term, and a definition. The analyst chooses the appropriate type (referring to the definition as needed) and enters it. (More precisely, the analyst enters the type's ID number and the term automatically appears in a portal.) If none of the known types matches the phenomenon under analysis, a new type is set up and defined and entered into the definitions file. There are presently 24 definition files, covering everything from the geographical definitions of subcontinental areas to kinds of morpho-syntactic phenomena. For instance, the definitions file for Locus (head/dependent marking) includes definitions for head, dependent, double, and zero marking as well as for less polar types: free or floating marking (e.g. in Wackernagel position); more than one head-marking formative on one word; dependent-marked word cliticized to head; various combinations; various splits; etc. The most recently added types are floating limited to specific lexemes as heads and the combination of head and floating marking. Locus information is entered in the module for marking of roles and the catalog of grammatical formatives. The definitions file for Position presently includes preposed, internal, postposed, circumposed, and Wackernagel position as well as five splits that have come up so far (e.g. some categories preposed and some postposed, as with Georgian subject-verb person-number agreement). Position information is entered on records for grammatical formatives and for verbal categories. The definitions file for Morphological Categories presently includes 52 records for the various types of inflectional categories so far found on verbs: e.g. tense, tense-aspect, tense-aspect-mood; Aktionsart; aspect; causative; etc. This file was set up with common types like person, number, gender, tense, role, etc. predefined, then others were added as they came up: categories such as Austronesian-style focus and Reciprocal/Reflexive were added soon afterwards; recently added are scope, motion, and rich construct. Other morphosyntactic definitions files include clause linkage types; classificatory vs. semantic inalienability; kinds of experiencer coding; type of morpheme (formative, syntactic word); part-of-speech and similar restrictions (nouns only, pronouns only, first and second persons, animate nouns, etc.); syntactic roles (A, S, U, etc.); degrees of fusion (isolating, concatenative, non-linear, reduplicating, etc.); kinds of flexivity (of stems or formatives: category-based, lexeme-based, both, neither). Each definition file is used in two or more different data modules. Opening a definitions file gives the researcher a complete taxonomy of the types attested so far, with definitions; the default sort by creation order gives a good sense of relative frequency; and actual frequency in the database can easily be checked by searching. Thus the process of data entry acquaints the researcher with the larger typology while contributing to that typology.

The data modules break descriptive notions down into their smallest units, and are unambiguous and operationalized. The ones we have so far cover several basic typological domains (e.g. locus, alignment) and some we have newly defined or refined: alienability (261 languages), covert categories (23 languages), exclusive/ inclusive (322 languages), grammatical markers (about 300 languages), locus per role (about 200 languages), morphological alignment (252 languages), NP structure (200+ languages), syntactic patterns (21 languages), synthesis (132 languages). Other modules are currently being developed (clause linkage, object downgrading, agreement types, etc.)

Finally, there are several backbone modules, designed as useful tools for the whole field: separate files for the language and its ID number and genetic classification; geographical location in terms of both area and coordinates; good-sized genetic and areal samples drawn from the larger database; and an EndNote™ bibliography of language data sources.

The Examplar-based Method

Among the descriptive parameters commonly used in modern reference grammars figure labels like degree of fusion ("agglutinative", "inflectional") or synthesis ("polysythentic"). Both authors and readers of grammars find these concepts useful as general descriptors although the concepts have been criticized for being ill-defined and for not having much predictive value in typology. The descriptive problem with fusion, synthesis and other such

parameters is that languages are often heterogenous: some part of a case paradigm may be highly fusional, other parts more agglutinative; likewise, verbs may look highly synthetic in some respects (e.g. allowing incorporation), but not in others (e.g. having only little inflectional morphology). Another problem arises when comparing, say, the position or locus of a given morphological category cross-linguistically. This is complicated by the fact that the content of categories differs so greatly that drawing a firm line around a category is usually impossible, and hence it is hard to be sure that one is comparing all and only comparabilia.

The AUTOTYP project overcomes these problems by using *exemplar-based* definitions for several categories whose structural treatment we compare cross-linguistically. Since descriptive grammars have to grapple with (or in some cases fail to grapple with) such matters as what to call tense, whether the language has tense, and whether the apparent lack of classic tense is worth mentioning, tracing an exemplar-based definition of an essential notion and looking at some applications of it in a small sample will be useful to description and useful to designers of questionnaires for standard descriptions.

Thus, for example, the AUTOTYP standard TAM category is defined as follows: "Generic inflectional tense, aspect, mood, status, etc. marker. If any of these markers differ from others in their morphological behavior, TAM refers to tense; if any tense marking differs from other tense markers, TAM refers to the tense used for basic, aspectually unmarked past time reference, i.e. to the form that serves as a simple response to 'what happened?'."

The AUTOTYP standard case category: "Generic inflectional case marker. If any of these markers differs from other in their morphological behavior, CASE refers to grammatical (core) case; if any core case marker differs from others, CASE refers to accusative, ergative or agentive case." In our survey (131 languages so far; Bickel & Nichols, in press), the following are some of the things that qualify as our standard case: the Belhare and Ingush ergative cases; the German accusative case; Russian differential object marking (accusative/ genitive); the Maori "preposition" /i/; the Japanese postposed particle /o/; the Cree obviative suffix; the Warao dative "postposition"; Yoruba tone sandhi on the object; the Spanish proclitic *a*; the Georgian dative case; the Squamish relative prefix. (Double quotes flag terms used in the literature that treat isolating grammatical formatives as though they were syntactic words.) Languages that lack any standard case representative include Songhai, Thai, Swahili, Slave, Abkhaz, Lakhota, and other languages that are radically head-marking in the clause.

The AUTOTYP standard noun plural: "Inflectional plural (or nonsingular) marking for nouns. If plural is different from dual, use the true plural. If different nouns have different morphological plural formations, use the commonest (or default) one. If animate (or human) and inanimate nouns have morphologically different plurals, enter both (creating two records for Noun Plural). (If there is yet another morphologically different salient non-default plural, that can also be entered in the database as well, but don't flag it as this plural.) If plural marking is described as optional, use it nonetheless (indicate optional nature in the database). Use only true plurals; do not use a collective, distributive, etc. if there is no plural."

Phenomena for which exemplar-based definitions have been used in previous literature include color terms, where cross-linguistic comparison uses focal color terms that are determined language by language; and word order, where comparison uses the word order of independent main clauses in pragmatically neutral contexts (and ignores special ordering rules in dependent clauses or all-new utterances etc.)

The coding method we use in AUTOTYP for verb synthesis is also examplar-based. Here, however, instead of sampling for one particular category, we sample for one particular maninfestation of synthesis: that degree of synthesis that is most extremely possible with verbal inflectional (not derivational) categories. That is, we take the categories and slots that can cooccur on the maximally inflected verb form. By *inflectional category* we understand any grammatical category whose presence or shape is (at least in part) a regular response to the grammatical environment. The prime candidates for this are categories like agreement, tense/aspect/mood, evidentials/miratives, status (realis, irrealis, etc.), polarity (negation), illocution (interrogative, declarative, imperative), and voice (including Austronesian-style verb orientation). Often, these categories are sensitive to the syntactic environment (e.g. argument NPs in the case of agreement, sequence of tense rules in the case of tense, cross-clausal anaphora in the case of voice, etc.). But often, the grammatical sensitivity is more narrowly morphological: different evidential or tense forms may imply different paradigms, or combine with different sets of aspect forms, or voices, etc.

Surveying inflectional categories on verbs produces lists of what can be synthetically expressed, and with how many morphological formatives. Some of the less well-known inflectional verb categories that were added to the list during data collection include: verb focus or emphasis, transitivity markers, construct marking (indicating the presence of a certain dependent NP, as in many African languages), object classifiers (inflectional if interacting with agreement as in Imonda), nonspecific reference-marking, scope (delimiting the scope of other categories), and causatives (inflectional when used in response to specific types of switch-reference patterns, as in Ingush).

Along with this in the AUTOTYP synthesis database also registers whether synthesis is or is not accompanied by regular incorporation (of nouns, verbs, adverbs etc.) and whether and to what degree synthetic grammatical words can be incoherent prosodically, morphophonologically or syntactically (incoherent in Dixon's 1977 sense; also cf. Hall, 1999).

The result is a scale of synthesis scores presently ranging from a low of 2 (Maybrat, Sango) to a high of 33 (Wichita) on which any language can be placed.

Thus exemplar-based definitions can make it possible to place one's field language in a cross-linguistic context at an early stage of descriptive work. This in turn will help identify hypotheses and priorities for further elicitation while the researcher is still in the field.

Conclusions

Fieldwork is in a sense "a way of catapulting oneself into the jungle language by the momentum of the home language" (Quine, 1960:70). The results are the better the less the home language is limited to any one particular grammar (e.g. the Scholastic grammar of Latin), and the more it is informed by typological knowledge of many different grammars: when applying analytic notions to a newly encountered structure, the researchers relies on what he or she knows about other languages. Typical questions arising from this are for example: is a given construction really a passive in the sense the term is used in familiar languages? If not, why not, and to what degree is it different? Is there need for another concept, or is this just a subspecies of a passive? etc. Judgment of these issues, and the resulting analysis, depend to a considerable degree on the fieldworker's knowledge of typological variation.

Dixon (1997) has proposed to call this typological knowledge *Basic Linguistic Theory*, thereby calling attention to the fact that this knowledge should not be seen as fieldworkers' personal expertise, but as an essential part of the linguistic sciences. Autotypologizing type inventories contribute to this part of our discpline. They have the advantage that they are constantly updated and are strictly inductive, i.e. not based on a priori models of possible and impossible human languages.

Use of an autotypologizing database as a reference will greatly increase the efficiency with which a field worker can incorporate theoretical and typological sophistication into a description. The former gap between theory and description is rapidly narrowing even without our help, but increased efficiency in gaining perspective on the typologically critical features in one's own field language will help field linguists, whose work is the most time-consuming of any in linguistics, economize time. We believe this may be especially useful in the early stages of fieldwork, when getting funding depends on demonstrating levels of both descriptive and comparative expertise that are hard to reach in the short time frame available for pilot studies. Autotypologizing together with exemplar-based definitions should also make fieldwork more efficient by increasing the likelihood that the researcher can identify and investigate a full range of relevant hypotheses while still in the field.

Acknowledgments

U.S. National Science Foundation Grant No. 96-16448 (Nichols, P.I.), Swiss National Science Foundation Grants No. 08210-053455 and 610-062717 (Bickel, P.I.), Institute for Slavic, Eurasian, and East European Studies, UC Berkeley.

References

Bickel, B. & Nichols, J. (in press). Fusion and exponence of selected formatives. In: Comrie, B., Dryer, M. S., Gill, D. & Haspelmath, M. (Eds.), *World atlas of linguistic structures*. Oxford: Oxford University Press.

Chelliah, S. L. (1997). *A grammar of Meithei.* Berlin: Mouton de Gruyter.

Conklin, H. (1955). Hanunóo color categories. *Southwest Journal of Anthropology* 11, 339 – 44.

Dixon, R. M. W. (1977). Some Phonological Rules of Yidiny. *Linguistic Inquiry* 8, 1 – 34.

Dixon, R. M. W. (1997). *The rise and fall of languages.* Cambridge: Cambridge University Press.

Hale, K. (1976). The adjoined relative clause in Australia. In: Dixon, R. M. W. (Ed.), *Grammatical categories in Australian languages.* (pp. 78 – 105). Canberra: Australian National University.

Hall, T. A. (1999). The phonological word: a review. In: Hall, T. A. & Kleinhenz, U. (Eds.), *Studies on the phonological word.* (pp. 1 – 22). Amsterdam: Benjamins.

Lucy, J. A. (1997). The linguistics of 'color'. In: Hardin, C. L. & Maffi, L. (Eds.), *Color categories in thought and language.* (pp. 320 – 46). Cambridge: Cambridge University Press.

Penchoen, T. G. (1973). *Tamazight of the Ayt Ndhir.* Los Angeles: Undena.

Quine, W. v. O. (1960). *Word and object.* Cambridge, Mass.: MIT Press.

Watkins, L. J. & McKenzie, P. (1984). *A grammar of Kiowa.* Lincoln: University of Nebraska Press.