Better Data with Late Aggregation: AUTOTYP and beyond

Balthasar Bickel[1], Taras Zakharko[1] and Johanna Nichols[2]

[1]University of Zurich, [2]University of California, Berkeley

balthasar.bickel@uzh.ch, taras.zakharko@uzh.ch, johanna@berkeley.edu

Contemporary typological databases often encode linguistic structure at a relatively coarse-grained level: structural variation in languages is aggregated into a single datapoint per language and fine-grained categories into broad types selected from a small list (e.g. 6 word orders, 3 primary morphosyntactic alignments). Following earlier work on bottom-up typologizing (Bickel & Nichols 2002), we develop here an alternative linguistic database design principle, Late Aggregation, and demonstrate its advantages and practical implementation with our work on the AUTOTYP database system over the past 20 years.

The core idea behind Late Aggregation is to postpone all high-level aggregation until the stage of data analysis (database use), well after data collection (database development). Specifically, we (a) encode the data using maximally fine-grained variables that are developed as the data collection progresses ('autotypologizing'), (b) use multiple datapoints per language when necessary, in order to adequately represent language-internal variation (e.g. capture word order per clause type rather than per language) and (c) use database modelling techniques to make the relationships between variables (analytic entities and their parameters) explicit (Chen 1976). When exploiting the database, variables can be aggregated into multiple types in parallel, and languages (or whole families) can be aggregated into single datapoints, in response to the research question at hand. For example, fine-grained coding of various head-marking variants can be aggregated in different ways, e.g. focusing on default vs. non-default treatments, or nouns vs. pronouns, or agreement vs. construct-state markers, or affix vs. clitic status. Similarly, language-internal variation can be aggregated variously into summary measures per language or variants that evolve in parallel.

Late Aggregation offers several substantial advantages:

(i) Sustainability — the same database can be reused for many different purposes. E.g. a fine-grained database on grammatical relation coding can be aggregated into a basic alignment typology (Bickel et al. 2014) or into a dataset for clustering semantic roles (Bickel et al. 2014).

(ii) Empirical responsibility — late aggregations are algorithmically defined and thus strictly derived from the underlying data. This brings typological work closer to primary grammatical description and improves its empirical responsibility by working with variables that are able to capture fairly detailed particulars across languages (Bickel 2010), reducing the risk that results are too heavily dependent on the definitions of high-level concepts (such as 'subordination').

(iii) Heuristics — premature aggregation limits the range of the signals and the range of relevant loci of variation that can be detected, e.g. we cannot anticipate the level of aggregation affected by language contact effects. Multiple parallel aggregations, coupled with false discovery rate assessments, avoids this problem (e.g. Bickel & Nichols 2006)

(iv) Durability — our database design promotes long-term currency and stability. Despite major changes in typological theory over the last 20 years, the AUTOTYP database is still used as a primary resource by many linguists. It has continued to grow, with new data points added almost daily, and is now the world's largest typological database (in terms of total datapoints) by an order of magnitude.

References:

Bickel, Balthasar. 2010. Capturing particulars and universals in clause linkage: a multivariate analysis. In Isabelle Bril (ed.), *Clause-hierarchy and clause-linking: the syntax and pragmatics interface*, 51–101. Amsterdam: Benjamins.

Bickel, Balthasar & Johanna Nichols. 2002. Autotypologizing databases and their use in fieldwork. In Peter Austin, Helen Dry & Peter Wittenburg (eds.), *Proceedings of the International LREC Workshop on Resources and Tools in Field Linguistics, Las Palmas, 26-27 May 2002*, Nijmegen: MPI for Psycholinguistics [http://www.autotyp.uzh.ch/download/canary.pdf].

Bickel, Balthasar & Johanna Nichols. 2006. Oceania, the Pacific Rim, and the theory of linguistic areas. *Proc. Berkeley Linguistics Society* 32. 3 – 15.

Bickel, Balthasar, Alena Witzlack-Makarevich & Taras Zakharko. 2014. Typological evidence against universal effects of referential scales on case alignment. In Ina Bornkessel-Schlesewsky, Andrej Malchukov & Marc Richards (eds.), *Scales: a cross-disciplinary perspective on referential hierarchies*, 7–43. Berlin: De Gruyter Mouton.

Bickel, Balthasar, Taras Zakharko, Lennart Bierkandt & Alena Witzlack-Makarevich. 2014. Semantic role clustering: an empirical assessments of semantic role types in non-default case assignment. *Studies in Language* 38. 485 – 511.

Chen, Peter. 1976. The entity-relationship model—toward a unified view of data. *ACM Transactions on Database Systems* 1, 1