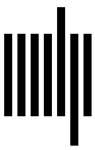
This is a section of doi:10.7551/mitpress/12200.001.0001

The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022



The MIT Press

56 Managing AUTOTYP Data: Design Principles and Implementation

Alena Witzlack-Makarevich, Johanna Nichols, Kristine A. Hildebrandt, Taras Zakharko, and Balthasar Bickel

1 Introduction

This data management use case describes AUTOTYP, a large-scale research program with goals in both quantitative and qualitative typology.[1] It was launched in 1996 by Balthasar Bickel and Johanna Nichols; however, individual data collection began much earlier. The theoretical framework was adopted in early 2001 and has been refined and elaborated on since then. The technological framework was adopted in 2001 as well and has experienced multiple updates since.

AUTOTYP is one of the oldest typological databases still in use and continuously developed for almost twenty-five years. Although the growth of the database proceeded uninterruptedly, several bursts of intensive data collection are associated with a number of research projects (see section 4).

The goals and principles of AUTOTYP follow from our understanding of the goals of linguistic typology more generally. They were originally formulated by Johanna Nichols in *Linguistic Diversity in Space and Time* (Nichols 1992). As such, we aim at identifying patterns in structural features among the world's languages—whether they are universal preferences or patterns with skewed distribution due to geographical or genealogical factors—and discovering principles governing their distribution. These goals are often phrased as searching for answers to the questions of "what's where why?" (Bickel 2007).

AUTOTYP is a typological database; that is, the kind of data it contains are generalizations about phonological and morphosyntactic structure of languages. The primary source of these generalizations is our interpretation of the analysis of annotated or structured forms of data collected from language use found in reference grammars or fieldwork (cf. Good's [chapter 3, this volume] discussion of the kinds of data used in the study of language in general, as well as more specifically in areal-typological studies).

The aim of this chapter is twofold. First, in section 2 we outline some fundamental design principles of AUTOTYP. Section 3 illustrates the implementation of these principles with one module of the AUTOTYP database, namely the grammatical relations module. This section will be of use to readers who plan to start their own typological database or who are already working on one. Section 4 is aimed at users and data consumers. It describes how to access the data and provides a brief overview of several use cases and research questions that have been asked with AUTOTYP.

2 The principles of AUTOTYP

Several other typological databases are comparable to AUTOTYP due to the nature of their goals and the type of data collected. However, from its first days, AUTOTYP followed a radically different design philosophy than the one adopted by many traditional typological databases, such as WALS (Dryer & Haspelmath 2013) or more recently Grambank (Grambank Consortium 2019). Whereas some of these principles are occasionally used by other databases, others remain truly unique to AUTOTYP. In what follows, we will first outline the five major principles of AUTOTYP: namely, modularity and connectivity (section 2.1), Autotypology (section 2.2), the division of labor between definition files and data files (section 2.3), the principle of late aggregation (section 2.4), and the exemplar-based method (section 2.5). Section 3 illustrates how these principles are implemented in practice using the example of the AUTOTYP module on grammatical relations.

2.1 Modularity and connectivity

AUTOTYP is not a single database but rather a network of thematically defined databases or modules on a wide variety of topics that all share the same infrastructure

and design principles. Each module can function as a stand-alone database, and it can easily be linked to other, already existing or future databases. This gives AUTOTYP databases the potential to grow in any possible direction without necessitating revisions of their basic design structures.

Most modules were developed by one or two researchers, often within the framework of a specific research grant. A data module typically covers a clearly defined typological domain with varying internal complexity ranging from relatively narrow ones with just a few variables in one or two tables (e.g., clusivity[2] or inflectional synthesis of the verb[3]) to broader and more complex ones with several dozen variables (e.g., noun phrase structure, clause linkage, word domains, grammatical markers). Structurally and conceptually, the most complex module is dedicated to grammatical relations: it encompasses several tables and over a hundred variables (see section 3).

We furthermore distinguish between service modules and data modules. The primary service module `Register`[4] contains an inventory of languages and their genealogical classifications and locations, as well as tables with linguistic areas, subsistence types, and sampling options. Initially, it also contained a bibliography module; however, now we keep track of our references using a BibTeX file.

All AUTOTYP modules (and the bibliography file) are linked together in a relational network via numerical language IDs, which are also mapped to other common language name IDs, such as Glottocodes from Glottolog (Hammarström, Forkel, & Haspelmath 2019), ISO 639-3 codes for the representation of names of languages, as well as WALS codes (Dryer & Haspelmath 2013). When developing a new module, we emphasize connectivity with existing modules and data reuse by different projects. The data module most other modules are connected to is the `Grammatical_markers` module. It contains information on over five thousand grammatical formatives (e.g., case and agreement markers, TAM affixes, and other grammatical markers) with details about their position, locus, degree of fusion, and exponence. Any other module—and these are the majority—that includes structures characterized by a specific grammatical formative is linked to this module.

2.2 Autotypology

Typological databases use categorical generalization as an abstraction tool. A closed, a priori determined list of possible values (i.e., an etic grid) is used to capture differing observations. Such lists can be motivated by tradition, for example, traditionally, case systems are classified as showing either accusative or ergative or neutral or tripartite or horizontal alignment (Comrie 1989:125–128). Idealized intuitions can also play a role; for example, a verb agreement system is classified as representing one of the five common alignment types or it is classified as a hierarchical agreement system, which is often represented as any system of agreement where a person hierarchy matters (e.g., Mallinson & Blake 1981). Theoretical considerations might as well be responsible. For instance, prior to the 1980s, in descriptive and theoretical accounts, clause linkage was believed to be either coordination or subordination. Later, another type—cosubordination—was added to this typology and became an integral part of some theoretical frameworks (Olson 1981; Foley & Van Valin 1984). Consequently, these three discrete categories show up in some typological surveys (e.g., Schmidtke-Bode 2009:150). However, as Bickel (2010) demonstrates, cross-linguistic variation in clause linkage is higher than what is allowed by these three categories, which entail sets of strictly correlated properties. Finally, in many cases, plain convenience is a crucial factor in categorical generalization. For instance, typologists who classify whole languages as belonging to one of the five morphological alignment types mentioned above are well aware of different types of splits and differential marking patterns (see, e.g., Comrie 2013a, 2013b on case alignment). However, it is simply easier to both code and process one-type-per-language data sets that do not require any further aggregations, as well as to do statistical analysis on one data point per language.

From its beginning, AUTOTYP chose a different approach, which we call the *autotypologizing method* in Bickel and Nichols (2002) or just *autotypology*. The idea is to prioritize coding adequacy and compatibility of the data with a wide range of theoretical frameworks over easy encoding with a list of predefined values. What does this shift in priorities mean in practice? Instead of sticking to an a priori defined etic grid, developers of AUTOTYP modules dynamically expand lists of possible values during data input (this characteristic is discussed

in more detail in Bickel & Nichols 2002). In practice, it looks as follows: Many modules indeed start with lists of possible values motivated by theoretical research and typological studies. However, we do not stop here: when coding a new language, we first check whether the previously established notions are sufficient for this language. If not, we postulate new possible values and carefully define them in a definition file (see section 2.3). This often means that, particularly at the beginning of a project, database entries are constantly revised, and the initial coding is replaced by more fine-grained values or even several variables, thus constantly reflecting the emerging typology of the phenomenon of interest (see section 3). This procedure is time-consuming in the beginning because the introduction of most new types requires a review and often a revision of all previous entries, but after about forty to fifty entries, new types become less likely to emerge and the typology stabilizes. The advantage of this procedure is data accuracy on a level that is impossible in databases with predefined typologies.

Occasionally, descriptive needs go beyond the mere revision of definition files. In this case, entire new variables (fields) or even tables are added to the respective module and all the data points entered to that point are reviewed and revised. As a general design principle, AUTOTYP favors the increase in the size and complexity of the database rather than in the complexity of coding decisions. For example, a database on clause linkage mentioned earlier does not adhere to the traditional distinction between coordination and (co)subordination and does not force each phenomenon in each coded language into one of these types; instead, it has gradually evolved into a set of specific variables that capture the full diversity of the phenomenon at hand.

2.3 Definition files versus data files

The Autotypology principles outlined in section 2.2 require differentiation between *data files* and *definition files*. Data files contain actual data on individual languages or constructions in individual languages. Definition files are essentially lists of possible values for each variable coded. In addition to category labels, they contain detailed linguistic definitions of each possible value, as well as a description of the coding procedure. As definition files are created dynamically and are updated throughout the whole process of data collection (see section 2.2), they thus reflect an empirically well-supported and detailed typology of the phenomenon at hand at any time. The two file types allow for dual use of the database in research: the data files allow quantitative typological inquiry into statistical correlations between structural, genealogical, or geographical features, while the definition files produce contributions to qualitative typology because they contain all and only notions that are cross-linguistically relevant and viable.

The binary distinction between data and definition files does not always work in practice. Some files have a dual status. For instance, the `Predicate_class_def` file serves as a definition file for the purposes of defining grammatical relations: it codes language-specific minor verb classes that have deviating coding patterns (see section 3). However, due to its language-specific character and the type of information coded (lists of predicates, type and token frequency of predicate classes, semantic domains, and so on) it can also be regarded as a data file and indeed it was used in Bickel et al. (2014) to answer the question whether there is cross-linguistic evidence for postulating clusters of predicate-specific semantic roles, such as experiencer, cognizer, or possessor.

2.4 Late aggregation

As we outlined in section 2.2, during data encoding, we choose the lowest-level, most exhaustive model that is appropriate to the data domain and the purpose of data collection. However, to answer specific research questions, the available data are typically filtered and aggregated. *Data aggregation* is any process in which information distributed over multiple values of a data set (or a subset of it) is grouped together and expressed in one single value, such as for purposes of statistical analysis. In linguistic typology, the most common variable for aggregation is probably aggregation by language. Data aggregation can apply simple mathematical functions. For instance, to calculate the degree of inflectional synthesis per language, every entry coding an inflectional category (e.g., polarity, evidentiality, argument role) that can be expressed in a synthetic word is counted as one and then all entries are added up. This yields one number per language that represents its degree of inflectional synthesis (see Bickel & Nichols 2013c for further details on this aggregation procedure). Other research questions require more elaborate aggregations and grouping by multiple variables.

In AUTOTYP, we systematically and explicitly adopt the principle of *late aggregation*. Late aggregation means

that no aggregation takes place at the stage of data collection. All aggregations are defined by algorithms applied to the data as collected and stored, outside the database. That is, the categories used during coding are not necessarily identical to the aggregated categories used in analyses. For instance, in the inflectional synthesis example we coded detailed information about every individual inflectional category. However, we did not code the degree of synthesis per se. This number was calculated outside the database.

One advantage of the late aggregation approach is sustainability: the same data can be reused to answer different research questions or to comply with different theoretical frameworks (see Buszard-Welcher, chapter 10, this volume). It also allows one to evaluate different but related generalizations simultaneously without the necessity for additional dedicated coding. Thus, a wide range of different and competing aggregations can be supported by the same data. The researcher—and not the data—controls the depth and the scope of aggregation. Furthermore, the late aggregation approach also provides for empirical responsibility: as the data encoding model is exhaustive (that is, no special cases or exceptions are left behind), it reduces chances of opaque mapping of language facts to possible values of a variable and the algorithmic form of aggregations allows tracing aggregated data points back to their original empirical basis. Finally, this design principle is durable: it ages well because the fine-grained underlying coding is typically less susceptible to shifts in theory and research questions than aggregated data points are.

2.5 Exemplar-based method

As we outline in section 2.2, during data encoding, we choose the lowest-level, most exhaustive model that is appropriate to the domain in question. While AUTOTYP allows one to record all this variation, for many typological surveys it is still desirable and more efficient to have one data point per language only. This can be achieved without early aggregation and without predefined lists of gross language types by following what we call the *exemplar-based method*: we select one particular exemplar of paradigms or structural domains as representative for the whole range of possible values or variables a language has to offer.

This exemplar is identified following a standard algorithmic definition. To answer many research questions, we aim at selecting a high-frequency or well-understood exemplar: for example, as exemplars for tense (in `Inflectional_synthesis`), we use tense morphology in general if all tense categories and markers have the same position and other properties; otherwise, we pick a synthetic basic present and non-imperfective past. As the exemplar for case marking alignment, we choose arguments of frequent verbs and independent clauses. Importantly, the choice of the exemplar is made not during the phase of data encoding but first at the stage of data aggregation. If desired, any other algorithmic definition for the identification of the exemplar can be adopted without having to recode the data. Also the exemplar-based method allows free addition of further data points when the need arises (or resources become available): one can simply add information of non-exemplar variants in each language, without any redesign of the database and its coding principles.

3 For developers: AUTOTYP principles at work

One of the most recent and most elaborate AUTOTYP modules is `Grammatical_relations`. In this section—primarily meant for developers of typological databases—we provide an overview of this module and highlight how the design principles outlined in section 2 were implemented in it.

The term *grammatical relations* (GRs) traditionally denotes the relations between a clause or a predicate and its arguments. The two traditional major types of GRs are subject and direct object. These categories are among the most basic concepts of many models. However, in response to many challenges with the way traditional GRs were identified and characterized (for a recent overview, see Witzlack-Makarevich 2019), recent, typologically supported research on GRs takes a construction-specific and language-specific view of GRs. That is, instead of adopting universal atomic notions of subject and object, one considers all relevant language-specific morphosyntactic properties of arguments (i.e., all relevant constructions) without prioritizing among them and without cherry-picking the ones that support the linguist's intuition. The general principles of this approach, as well as major variables were first outlined in Bickel (2011). The manuscript of that paper served as the starting point for the development of the AUTOTYP module on grammatical relations in 2006.

GRs are defined as equivalence sets of arguments, treated the same way by some construction (or *argument selector*) in a language (Bickel 2011). Cross-linguistically common argument selectors are, for example, case marking, agreement on the verb, or passivization. Languages vary in terms of how many argument selectors they have (see, e.g., the collection of papers in Witzlack-Makarevich & Bickel 2019).

The data in the module on GRs are extracted from primary source documents (grammars and articles), and only occasionally are they obtained via personal communication from speakers and specialists. The extraction of the necessary data is first recorded in a language report: a text document that apart from the decision on the assignment of values and motivation behind it has multiple examples, paradigms, and citations from the primary sources. The language reports are particularly useful at the initial stages of module development when new values are identified and added to the value lists and coding decisions need to be revised (see section 2.2). They also prove to be useful when further variables are added. For instance, the GRs report often contains full paradigms of verb agreement. Originally, we only coded for a high-level agreement variable (i.e., which argument the verb agrees with), at a later point we wanted to expand the data set with details about overt and zero agreement markers, as well as portmanteau markers. In this case, we relied on the paradigm in the language reports and only occasionally had to consult primary sources.

Being a relational database, the `Grammatical_relations` module can be understood as a collection of relations, which are perceived by the user as related tables (for an introduction to relational databases, see, e.g., Harrington 2016 or Kroenke et al. 2019). Each table is a set of data elements (or values) in the form of rows and columns. Each row (or record) corresponds to some object (e.g., a selected argument, a grammatical relation, or a language). Each column (also called field or attribute) represents a property of this object. The linking between individual files is realized by means of a common field (an identifier). To relate any two files, they simply need to have such a common field.

In line with the definition given in the beginning of this section, the major entity of the `Grammatical_relations` module is a single GR coded in the `Grammatical_relation` table illustrated in table 56.1. Every record in the `Grammatical_relation` table contains information about the language in question (linked to the `Language` table of the `Register` module via a unique language identifier [`LID`]), the argument selector (e.g., a specific case) that forms the subset, and the selected arguments. Thus, each GR of a language as defined above forms one record in the database. For every language there are as many records as there are GRs established by various argument selectors. For instance, in the example from Hindi in table 56.1, there are separate entries for individual cases, as well as entries for agreement and syntactic constructions, such as raising to object. This aspect differentiates the module from those typological databases that have individual languages as central entities. As of February 2020, the module contains data on 4,400 argument selectors in 779 languages.

The data entry is done via various layouts in FileMaker Pro specifically developed for the purposes of data entry. One of the layouts is shown in figure 56.1. Many similar layouts for data entry are developed on the fly to make individual data entry tasks easier.

The field `SourceID` contains BibTeX keys that link this database to our bibliography database. We keep track of our references using BibTeX, a reference management software.

Table 56.1
An example of the `Grammatical_relation` table with a selection of entries for Hindi (`LID` 99)

GRID	LID	SelectorID	Traditional_term	Grammatical_markerID	Selected_itemID	SourceID
63	99	2	Genitive case	32	214, 215, 551, 552, . . .	Mohanan1994Argument
92	99	2	Ergative case	29	201, 202, 205, 209, . . .	Montaut2004Grammar
94	99	2	Nominative case	28	195, 196, 197, 198, . . .	Montaut2004Grammar
93	99	2	Dative case	30	126, 178, 312, 311, . . .	Montaut2004Grammar
2881	99	2	Locative case	33	21636, 21637	Montaut2004Grammar
97	99	3	Agreement (trigger potential)	31	125, 461, 462, 464, . . .	Montaut2004Grammar
56	99	7	Raising to object	NA	543, 544, 455, 546, . . .	Bickeletal2000Fresh

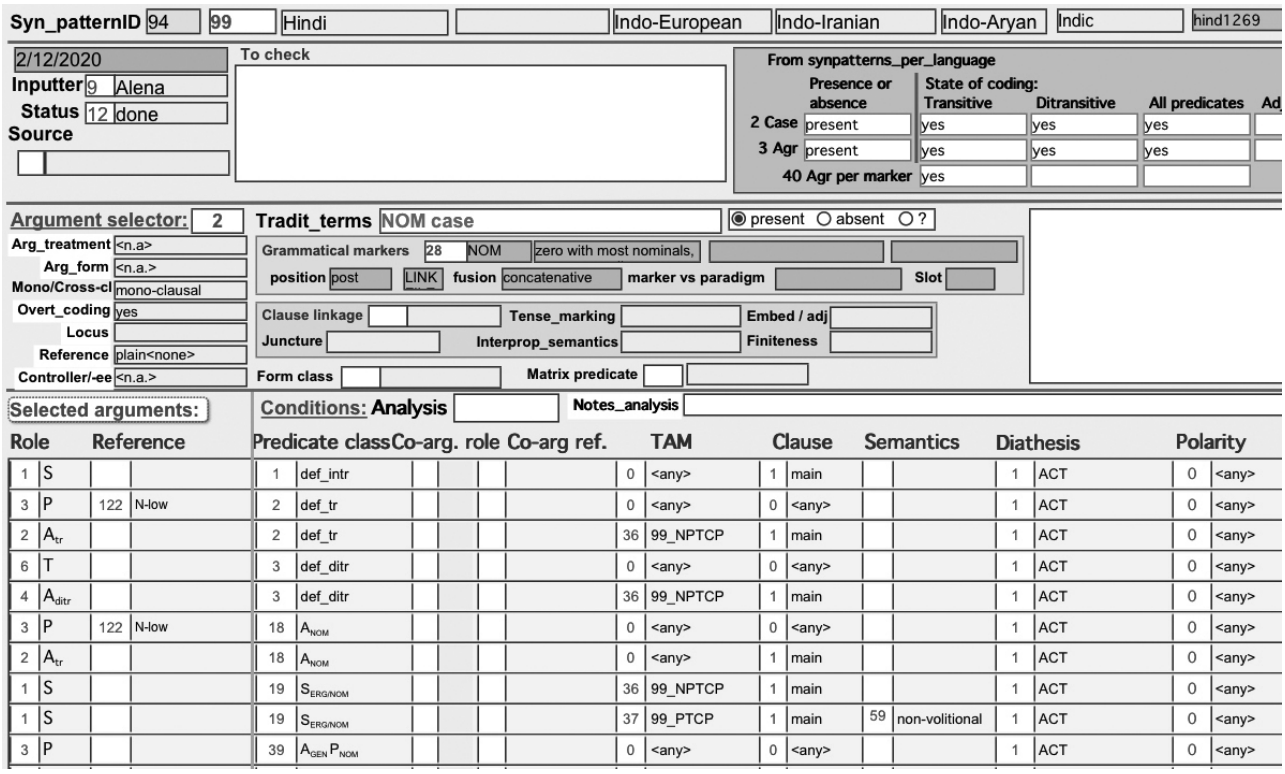


Figure 56.1
One of the FileMaker Pro interfaces of the AUTOTYP `Grammatical_relations` module.

BibTeX is stored in a plain text UTF-8 file. An example of the BibTeX entry for the first line in table 56.1 follows:

```
@book{Mohanan1994Argument,
  Address={Stanford, CA},
  Author={Mohanan, Tara},
  Publisher={Center for the Study of Language and
  Information},
  Title={Argument structure in Hindi},
  Year={1994}}
```

The `Grammatical_relation` table is linked to a number of tables, schematically represented in figure 56.2. First, it is linked to the `Language` table from the `Register` module via `LID`. This table provides the language name as used in AUTOTYP (e.g., Hindi) and alternative names, the genealogical affiliation, the area where the language is used, as well as language IDs from Glottolog (Hammarström, Forkel, & Haspelmath 2019, `ind1269` for Hindi) and ISO 639-3 (`hin`) for easy mapping to other databases.

The implementation of the AUTOTYP principle of modularity (section 2.1) can be illustrated with the modules `Grammatical_relations` and `Register`. The general information about individual languages (3,012 entries as of February 2020) is stored in the `Language` table of the `Register` module. It contains the information on the genealogy (e.g., branch and stock names) and geographic distribution (e.g., areas and coordinates) of individual languages. The primary key is a numerical language ID (`LID`). `LID`s are used to link the general information about languages to files on various aspects of grammar, among them to the `Grammatical_relations` module. In turn, as we have mentioned, `LID`s are associated with other codes, for instance, Glottolog (Hammarström, Forkel, & Haspelmath 2019) and ISO 639-3. This allows for a straightforward compatibility of the AUTOTYP databases with other databases following one of these standards.

As figure 56.2 shows, the `Grammatical_relation` table is in a many-to-one relationship with the `Selector` table, which specifies the precise nature of the argument selector and contains such variables as whether the selector is a coding construction (e.g., agreement) or a behavior construction (e.g., control of reference), whether it is a mono-clausal or cross-clausal construction, whether it involves head or dependent marking, and so forth (see Witzlack-Makarevich 2011 for details on the typology of argument selectors).

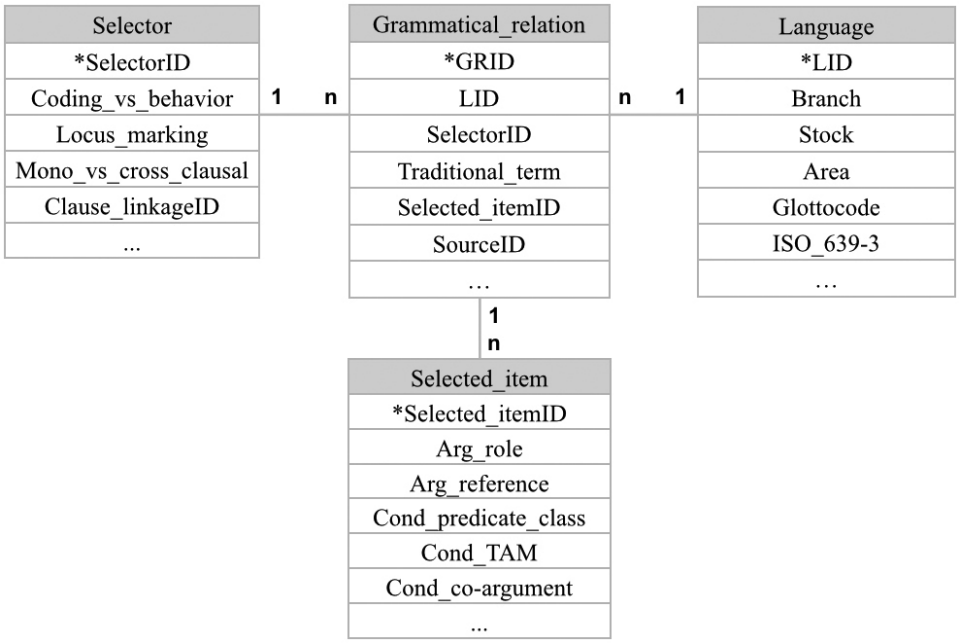


Figure 56.2
`Grammatical_relations` module in AUTOTYP.

Table 56.2
An example of the `Selected_item` table in AUTOTYP listing a selection of entries for the nominative case in Hindi

Selected_itemID	Arg_roleID	Arg_referenceID	Cond_predicate_classID	Cond_TAMID	Cond_clauseID
195	1 (=S)	0 (=any)	1 (=default_mono)	0 (=any)	1 (=main)
196	2 (=A)	0 (=any)	2 (=default_bi)	36 (=99_NPTCP)	1 (=main)
198	3 (=P)	122 (=N-low)	2 (=default_bi)	0 (=any)	1 (=main)
197	4 (=A_{ditr})	0 (=any)	3 (=default_tri)	36 (=99_NPTCP)	1 (=main)
203	6 (=T)	0 (=any)	3 (=default_tri)	0 (=any)	1 (=main)
237	2 (=A)	0 (=any)	18 (=A_{NOM})	0 (=any)	1 (=main)
4172	3 (=P)	122 (=N-low)	18 (=A_{NOM})	0 (=any)	1 (=main)
13804	3 (=P)	0 (=any)	39 (=A_{GEN} P_{NOM})	0 (=any)	1 (=main)

Following the principle of the differentiation between data files and definition files (see section 2.3), for each argument selector, `Selected_item` lists only the IDs for role and reference of selected arguments and all relevant conditions on argument selection (e.g., predicate class, TAM, type of clause). The exact specifications of the values behind these IDs are stored in the respective definition files. However, for friendliness to human readers, table 56.2 includes in parentheses the labels of the individual values from the respective definition files. (For our definition of the argument roles S, A, P, A_{ditr}, T, and G, see Bickel 2011 and Witzlack-Makarevich 2019; NPTCP stands for a non-participle verb form.)

A typical definition file is illustrated with a few entries in table 56.3. For instance, it spells out what is behind the `Arg_referenceID` 122 from table 56.2.

A less typical and more complex definition file is illustrated in table 56.4. This is the `Predicate_class` table, which captures predicate classes as condition on argument selection, in other words, the situation where a specific case marking or construction is available only to arguments of a specific group of predicates. For instance, the majority of Hindi A arguments vary between nominative and ergative case marking as conditioned by the morphological form of the predicate, such as whether it involve a participle or not—a situation often confused with TAM-based split in marking. However, a few Hindi predicates (`Predicate_classID` 18) do not participate in this alternation: their A argument is always in the nominative no matter what the shape of the predicate is. The definition files represent a taxonomy of encountered typological types and thus feed qualitative typology.

Table 56.3
An example of the `Reference` definition table in AUTOTYP

ReferenceID	Label	Description
0	Any	
5	Pro	Free pronouns that head NPs, not pronominal agreement markers
105	1sgPro	First-person singular pronoun
123	3duPro	Third-person dual pronoun
34	N-anim	Animate noun
121	N-high	Noun with a higher discourse rank than "N-low" (where rank is determined by discourse factors with language-specific weights)
122	N-low	Noun with a lower discourse rank than "N-high" (where rank is determined by discourse factors with language-specific weights)

Neither table 56.1 nor table 56.2 contains any explicit information on the alignment of the case marking (e.g., "S=A≠P" or "nominative-accusative alignment"); however, this is often the kind of information researchers are interested in when studying grammatical relations. AUTOTYP differs from traditional typological databases in that in most cases, data are entered in a fairly raw format, as in table 56.1. It is comparable to reference grammar descriptions that list how e.g. individual cases, such as the nominative case, are used, i.e. which argument roles they cover and under which conditions they occur. For most analytical purposes, these data will be filtered, aggregated, and reshaped.

Thus, we systematically follow the principle of late aggregation (see section 2.4) and avoid any data aggregation at the stage of data collection. All aggregations are defined algorithmically in R (or occasionally Python) scripts outside the database. For instance, for the analysis presented in Bickel, Witzlack-Makarevich, Choudhary, et al. (2015) we only needed the data on the degree of ergativity of case marking. For this purpose, we first filtered from the `Grammatical_relation` table (table 56.1) only the entries on case marking and ignored all other argument selectors. Then, only a subset of entries with the relevant `selected_itemIDs` was filtered from the `Selected_item` table (table 56.2). A range of further filtering conditions were imposed to select the desired exemplar (see section 2.5 on the exemplar-based method), for example, we were only interested in the case marking of the arguments of the default mono- and bivalent predicate classes and only in active clauses. After the data were filtered properly, we proceeded with the aggregation and specifically considered whether the A argument is marked differently from the S argument (the case marking of the P argument was irrelevant for this research question). As some languages have split case marking, so that parts of the system align S and A and parts of the system do not, some languages have multiple entries. The aggregated results for Hindi are shown in table 56.5, and they can be further aggregated to one quantified alignment statement per language, for example, Hindi is to 25% ergative (S≠A). A range of other filter options and aggregations was performed on the same data (the relevant references can be found in section 4).

4 For users: Overview and use cases

There has been a broad range of published research based on the AUTOTYP data. A number of early data sets from AUTOTYP were integrated into WALS (Haspelmath et al. 2005) and are accessible via the WALS website (https://wals.info/; Dryer & Haspelmath 2013). These data sets are aggregations from the modules on the locus of marking (i.e., head vs. dependent marking, Nichols & Bickel 2013a, 2013b, 2013c), exponence and fusion of selected inflectional formatives from the module on grammatical markers (Bickel & Nichols 2013a, 2013b), and inflectional synthesis of the verb (Bickel & Nichols 2013c) from the module on verb synthesis. The noun phrase structure module provided the data for two further aggregated WALS chapters, Bickel and Nichols (2013d) on obligatory possessive inflection and Nichols and Bickel (2013d) on possessive classification, as well as for a PhD thesis by Rießler (2011). Another early data set comprises data on clusivity and is analyzed in Bickel and Nichols (2005). Data sets from the module on word domains served as the empirical base for aggregations in Bickel, Hildebrandt, and Schiering (2009) and Schiering, Bickel, and Hildebrandt (2010, 2012). The data set used to develop the module on clause linkage is aggregated and analyzed in Bickel (2010).

In recent years our own research has produced a number of publications based on various subsets of the data from the module on grammatical relations. The theoretical foundation behind this module was outlined in Bickel (2011). Its structure is described in detail

Table 56.4
An example of the `Predicate_class` definition table in AUTOTYP with some default and Hindi predicate classes

Predicate_classID	Label	Description	Translation_equivalent	Class_size	LID
1	Default monovalent	Monovalent predicate class with the default (or canonical) marking pattern of behavior	(Open class)	Large	0 (=any)
2	Default bivalent	Bivalent predicate class with the default (or canonical) marking pattern of behavior	(Open class)	Large	0 (=any)
3	Default trivalent	Trivalent predicate class with the default (or canonical) marking pattern of behavior	(Open class)	Large	0 (=any)
18	A_{NOM}	Bivalent Hindi-specific predicate class whose A argument is always in the nominative case and never in the ergative case	116 (=bring), 117 (=forget)	Very small (<5)	99 (=Hindi)
39	A_{GEN} P_{NOM}	Bivalent Hindi-specific predicate class whose A argument is in the genitive case and P argument in the nominative case	136 (=have)	Very small (<5)	99 (=Hindi)

Table 56.5
An example of data aggregation on the basis of the data in the `Grammatical_relations` module of AUTOTYP

LID	S_A_alignment	Clause_type	Structural_condition
99 (=Hindi)	S = A	Dependent	99_NPTCP
99 (=Hindi)	S = A	Main	99_NPTCP
99 (=Hindi)	S = A	Dependent	99_PTCP
99 (=Hindi)	S ≠ A	Main	99_PTCP

in Witzlack-Makarevich (2011). A subset of the database capturing the proportion of ergative alignment of case marking in some six hundred languages was used as the basis for our claim that languages tend to avoid ergatives when they evolve over time (Bickel, Witzlack-Makarevich, Choudhary, et al. 2015). A different subset of the data on case marking was used to test the hypothesis that if a language has differential subject or differential object marking the distribution of cases results from a universal effect of referential scales (first tested in Bickel & Witzlack-Makarevich 2008 and then on a larger data set in Bickel, Witzlack-Makarevich, & Zakharko 2015). Various hypotheses related to the principles underlying the subject and object agreement systems on the verb were tested using the agreement subset of the data in Bickel et al. (2013), Bickel, Witzlack-Makarevich, Zakharko, and Iemmolo (2015), and Witzlack-Makarevich et al. (2016). In each case a slightly different aggregation of the data was performed. In contrast to other publications based on the subset of case and agreement marking with major verb classes, Bickel et al. (2014) considers the subset of the data dedicated to case marking in minor verb classes in over 140 languages.

Starting with the very first AUTOTYP publications we prioritized openness, both in terms of transparency of methodology, as well as in terms of data accessibility (see Gawne & Styles, chapter 2, this volume, on open research in the social sciences). For many of the papers listed herein, the aggregation scripts, aggregated data, as well as scripts used to perform the statistical analysis are available via the publishers' websites, whenever the publishers provided this option. For instance, the data used in Bickel et al. (2014) are available as an online appendix at https://doi.org/10.1075/sl.38.3.03bic.additional. The data used in Bickel, Witzlack-Makarevich, Choudhary, et al. (2015) include the database of case-marking patterns (a `.csv` file, https://doi.org/10.1371/journal.pone.0132819.s001), bibliographical references of the sources

used (a BibTeX file, https://doi.org/10.1371/journal.pone.0132819.s002), as well as an R script with step-by-step results of the language evolution analysis (https://doi.org/10.1371/journal.pone.0132819.s003). Small data sets were occasionally added directly as tables in an appendix, as in Bickel (2010) or Schiering, Bickel, and Hildebrandt (2012). The data sets mentioned above as well as other data sets have been made available via AUTOTYP website at https://github.com/autotyp/autotyp-data.

In addition to the many data sets made available on various platforms over the course of the last twenty years, Bickel et al. (2017) is the first major release of over thirty AUTOTYP data sets (in over fifty tables) accompanied by metadata files. This release includes over one thousand variables with a total of about 4.5 million typological data points. We use `.csv` format for the data, `.yaml` format for metadata, and `.bib` format for bibliographical references. In addition, the geographical data is available in `.kml` format (see Mattern [chapter 5, this volume] on sustainable data formats and Han [chapter 6, this volume] on data formats and conversion). Finally, we also provide the entire data set as a list in R's `.rds` format. We have archived this release on Zenodo (see Andreassen [chapter 7, this volume] for a discussion of issues of data archiving). Zenodo provides a new digital object identifier (DOI) for each new release along with a bibliographic citation that users can use when citing a particular release of the data in their research papers.

In general, AUTOTYP data sets have rarely been used by researchers who were not part of the project. We believe that the major hurdle is that typologists still almost universally operate under the premise that whole languages are the proper level on which to code data. In our experience, the importance of this premise declines to the extent that researchers learn how to write algorithms for data aggregation. The current move toward training linguists in methods of modern data science, including basic scripting techniques and statistics, makes us confident that future generations will overcome the traditional hurdle of adopting AUTOTYP principles and using AUTOTYP data. An example of this is the recent use of AUTOTYP data by Schmidtke-Bode and Levshina (2018), who challenge the results of Bickel, Witzlack-Makarevich, and Zakharko (2015) by performing alternative statistical analyses.

Notes

1. We are grateful to Andrea L. Berez-Kroeker, Steven Moran, and an anonymous reviewer for quick and helpful comments on an earlier version of this chapter.

2. The data set is available as part of Bickel et al. (2017) under https://github.com/autotyp/autotyp-data/blob/master/data/Clusivity.csv.

3. The data set is available as part of Bickel et al. (2017) under https://github.com/autotyp/autotyp-data/blob/master/data/Synthesis.csv.

4. We use the monospaced font to refer to AUTOTYP's modules, tables, or fields, as well as to file extensions. Starting from Bickel et al. (2017), we consistently use upper camel case for variable names. We capitalize the first letter and use underscore in table and module names. For data entry, we used FileMaker Pro—a cross-platform relational database application from Claris International, a subsidiary of Apple Inc. It integrates a database engine with a graphical user interface (GUI). FileMaker Pro databases, which correspond to our modules, have the extension `.fmp12`. Individual tables correspond to files in these databases and as such are not individual files, for this reason no extension is given when we refer to them in the rest of the chapter. They can be exported in various formats (e.g., `.csv` or `.tab`), as discussed in section 4.

References

Bickel, Balthasar. 2007. Typology in the 21st century: Major current developments. *Linguistic Typology* 11 (1): 239–251.

Bickel, Balthasar. 2010. Capturing particulars and universals in clause linkage: A multivariate analysis. In *Clause-Hierarchy and Clause-Linking: The Syntax and Pragmatics Interface*, ed. Isabelle Bril, 51–102. Amsterdam: John Benjamins.

Bickel, Balthasar. 2011. Grammatical relations typology. In *The Oxford Handbook of Language Typology*, ed. Jae Jung Song, 399–444. Oxford: Oxford University Press.

Bickel, Balthasar, Kristine A. Hildebrandt, and René Schiering. 2009. The distribution of phonological word domains: A probabilistic typology. In *Phonological Domains: Universals and Deviations*, ed. Janet Grijzenhout and Kabak Barış, 47–75. Berlin: Mouton de Gruyter.

Bickel, Balthasar, Giorgio Iemmolo, Taras Zakharko, and Alena Witzlack-Makarevich. 2013. Patterns of alignment in verb agreement. In *Languages across Boundaries: Studies in the Memory of Anna Siewierska*, ed. Dik Bakker and Martin Haspelmath, 15–36. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110331127.15.

Bickel, Balthasar, and Johanna Nichols. 2002. Autotypologizing databases and their use in fieldwork. In *Proceedings of the*

International LREC Workshop on Resources and Tools in Field Linguistics, Las Palmas, 26–27 May 2002, ed. Austin, Peter, Helen Dry, and Peter Witternburg. Nijmegen: ISLE and DOBES.

Bickel, Balthasar, and Johanna Nichols. 2005. Inclusive-exclusive as person vs. number categories worldwide. In *Clusivity: Typology and Case Studies of the Inclusive–Exclusive Distinction*, ed. Elena Filimonova, 49–72. Amsterdam: John Benjamins.

Bickel, Balthasar, and Johanna Nichols. 2013a. Exponence of selected inflectional formatives. In *The World Atlas of Language Structures Online*, ed. Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/21.

Bickel, Balthasar, and Johanna Nichols. 2013b. Fusion of selected inflectional formatives. In *The World Atlas of Language Structures Online*, ed. Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/20.

Bickel, Balthasar, and Johanna Nichols. 2013c. Inflectional synthesis of the verb. In *The World Atlas of Language Structures Online*, ed. Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/22.

Bickel, Balthasar, and Johanna Nichols. 2013d. Obligatory possessive inflection. In *The World Atlas of Language Structures Online*, ed. Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/58.

Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga, and John B. Lowe. 2017. *The AUTOTYP Typological Databases*. Version 0.1.0 https://github.com/autotyp/autotyp-data/tree/0.1.0. http://doi.org/10.5281/zenodo.3667562.

Bickel, Balthasar, and Alena Witzlack-Makarevich. 2008. Referential scales and case alignment: Reviewing the typological evidence. In *Scales*, ed. Andrej Malchukov and Marc Richards (=Band 86 der Linguistischen ArbeitsBerichte). Leipzig: Institut für Linguistik.

Bickel, Balthasar, Alena Witzlack-Makarevich, Kamal K. Choudhary, Matthias Schlesewsky, and Ina Bornkessel-Schlesewsky. 2015. The neurophysiology of language processing shapes the evolution of grammar: Evidence from case marking. *PLOS ONE* 10 (8): e0132819. doi:10.1371/journal.pone.0132819.

Bickel, Balthasar, Alena Witzlack-Makarevich, and Taras Zakharko. 2015. Typological evidence against universal effects of referential scales on case alignment. In *Scales: A Cross-Disciplinary Perspective on Referential Hierarchies*, ed. Ina Bornkessel-Schlesewsky, Andrej Malchukov, and Marc Richards, 7–44. Berlin: De Gruyter Mouton.

Bickel, Balthasar, Alena Witzlack-Makarevich, Taras Zakharko, and Giorgio Iemmolo. 2015. Exploring diachronic universals of agreement: Alignment patterns and zero marking across person categories. In *Agreement from a Diachronic Perspective*, ed. Jürg Fleischer, Elisabeth Rieken, and Paul Widmer, 29–52. Berlin: De Gruyter Mouton.

Bickel, Balthasar, Taras Zakharko, Lennart Bierkandt, and Alena Witzlack-Makarevich. 2014. Semantic role clustering: An empirical assessment of semantic role types. *Studies in Language* 38 (3): 485–511.

Comrie, Bernard. 1989. *Language Universals and Linguistic Typology: Syntax and Morphology*. 2nd ed. Chicago: University of Chicago Press.

Comrie, Bernard. 2013a. Alignment of case marking of full noun phrases. In *The World Atlas of Language Structures Online*, ed. Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/98.

Comrie, Bernard. 2013b. Alignment of case marking of pronouns. In *The World Atlas of Language Structures Online*, ed. Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/99.

Dryer, Matthew S., and Martin Haspelmath, eds. 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/.

Foley, William A., and Robert D. Van Valin Jr. 1984. *Functional Syntax and Universal Grammar*. Cambridge: Cambridge University Press.

Grambank Consortium, eds. 2019. *Grambank*. Jena: Max Planck Institute for the Science of Human History. http://grambank.clld.org.

Hammarström, Harald, Robert Forkel, and Martin Haspelmath. 2019. *Glottolog 4.1*. Jena, Germany: Max Planck Institute for the Science of Human History. http://glottolog.org.

Harrington, Jan L. 2016. *Relational Database Design and Implementation*. 4th ed. Burlington, MA: Morgan Kaufmann.

Haspelmath, Martin, Matthew S. Dryer, David Gil, and Bernard Comrie. 2005. *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Kroenke, David M., David Auer, Scott L. Vandenberg, and Robert C. Yoder. 2019. *Database Concepts*. 9th ed. New York: Pearson.

Mallinson, Graham, and Barry Blake. 1981. *Language Typology. Cross-Linguistic Studies in Syntax*. Amsterdam: North-Holland.

Nichols, Johanna. 1992. *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Nichols, Johanna, and Balthasar Bickel. 2013a. Locus of marking in possessive noun phrases. In *The World Atlas of Language Structures Online*, ed. Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/24.

Nichols, Johanna, and Balthasar Bickel. 2013b. Locus of marking in the clause. In *The World Atlas of Language Structures Online*, ed. Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/23.

Nichols, Johanna, and Balthasar Bickel. 2013c. Locus of marking: Whole-language typology. In *The World Atlas of Language Structures Online*, ed. Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/25.

Nichols, Johanna, and Balthasar Bickel. 2013d. Possessive classification. In *The World Atlas of Language Structures Online*, ed. Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/59.

Olson, Michael L. 1981. Barai clause juncture: Toward a functional theory of inter-clausal relations. PhD dissertation, Australian National University.

Rießler, Michael. 2011. Typology and evolution of adjective attribution marking in the languages of Northern Eurasia. PhD dissertation, University of Leipzig.

Schiering, René, Balthasar Bickel, and Kristine A. Hildebrandt. 2010. The prosodic word is not universal, but emergent. *Journal of Linguistics* 46 (3): 657–709.

Schiering, René, Balthasar Bickel, and Kristine A. Hildebrandt. 2012. Stress-timed = word-based? Testing a hypothesis in prosodic typology. *STUF—Language Typology and Universals* 65 (2): 157–168.

Schmidtke-Bode, Karsten. 2009. *A Typology of Purpose Clauses*. Amsterdam: John Benjamins.

Schmidtke-Bode, Karsten, and Natalia Levshina. 2018. Reassessing scale effects on differential case marking: Methodological, conceptual and theoretical issues in the quest for a universal. In *Diachrony of Differential Argument Marking*, ed. Ilja A. Seržant and Alena Witzlack-Makarevich, 509–537. Berlin: Language Science Press.

Witzlack-Makarevich, Alena. 2011. Typological variations in grammatical relations. PhD dissertation, University of Leipzig.

Witzlack-Makarevich, Alena. 2019. Argument selectors. A new perspective on grammatical relations: An introduction. In *Argument Selectors: A New Perspective on Grammatical Relations*, ed. Alena Witzlack-Makarevich and Balthasar Bickel, 1–38. Amsterdam: John Benjamins.

Witzlack-Makarevich, Alena, and Balthasar Bickel, eds. 2019. *Argument Selectors: A New Perspective on Grammatical Relations*. Amsterdam: John Benjamins.

Witzlack-Makarevich, Alena, Taras Zakharko, Lennart Bierkandt, Fernando Zúñiga, and Balthasar Bickel. 2016. Decomposing hierarchical alignment: Co-arguments as conditions on alignment and the limits of referential hierarchies as explanations in verb agreement. *Linguistics* 54 (3): 531–561. https://doi.org/10.1515/ling-2016-0011.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363