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
1 Introduction

Studies in linguistic typology may pursue various types of epistemological objectives. As the term *typology* suggests, early approaches, inspired by structuralist biology, aimed at taxonomizing languages—not in genealogical or geographical terms, but with respect to structural properties (e.g., Schlegel’s 1808 morphological types), often with the idea of identifying the (holistic) “character” of a language (von der Gabelentz 1891). In the twentieth century, the “Greenbergian” approach to typology became prevalent (Greenberg 1966; Comrie 1989; Croft 2003), seeking to identify correlations between properties of languages and, ultimately, finding “linguistic universals.” Toward the end of the twentieth century, a “distributional” approach to linguistic typology established itself (Nichols 1992, 2007), seeking to determine—and explain—the distribution of linguistic features in time and space (“What’s where why?,” cf. Bickel 2007, 2015). This approach is also associated with the idea of focusing on more fine-grained properties of linguistic systems, making linguistic typology a multivariate endeavor (cf. Bickel 2010). What all major paradigms within linguistic typology have in common is that they deal with linguistic variation and the limits of that variation.

The approaches to linguistic typology mentioned have traditionally been based on generalizations made in grammatical descriptions and can thus be subsumed under the term *grammar-based typology* (though studies in lexical typology are based on dictionaries or word lists rather than grammars). More recently, attempts have been made to carry out typological studies on the basis of textual data, thus establishing *corpus-based typology* (cf. Goldhahn, Quasthoff, & Heyer 2014; Futrell, Mahowald, & Gibson 2015; Culbertson 2017; Alzetta et al. 2018). Corpus-based approaches mostly address the questions of Greenbergian typology, but the use of corpus data obviously opens up new possibilities in terms of statistical modeling.

The present chapter provides an overview of central matters of research data management for major approaches to linguistic typology as we have characterized. It focuses on grammar-based research, but the data model described in section 3 can, mutatis mutandis, be applied to corpus-based approaches as well. Following this brief introduction, section 2 addresses some general questions concerning empirical approaches to linguistic typology. Section 3 provides a description of one particular typological database, the Typological Database of Impersonals (ImproType). Section 4 contains some remarks on data storage and export, as well as examples of research questions addressed using data from ImproType. Some concluding remarks are made in section 5.

2 Empirical approaches to linguistic typology

As Chomskyan universalism is broadly rejected in linguistic typology, the question arises how the elements underlying crosslinguistic comparison—functional domains (e.g., future tense, impersonalization), linguistic categories (e.g., adjective, affix), and relations holding between members of these categories (e.g., subject, agreement)—can be defined in such a way that comparability can be assumed. The present chapter does not focus on this question, which has been widely debated without a general consensus being reached. Haspelmath (2010:664) introduced the term *comparative concept* as a solution to this problem: “concepts specifically designed for the purpose of comparison that are independent of descriptive categories” (664). For example, typologists have long been aware that there is no crosslinguistically applicable notion of “subject” (e.g., Keenan 1976), so a
statement such as “in language L, the subject S of a verbal predicate V generally precedes V” (i.e., “the language is SV”) is impossible to verify or falsify. According to Haspelmath’s approach, linguists can define a comparative concept “subject,” for example, using criteria like the ones discussed by Keenan.

As comparative concepts are theoretical concepts and thus not directly observable, they can only be defined intentionally. To be used in a crosslinguistic study, they consequently have to be operationalized. Operationalization is “the construction of actual, concrete measurement techniques” for theoretical constructs (Babbie 1989:5). Consequently, a procedure has to be devised that allows researchers to identify the (theoretical, intentionally defined) comparative concept “subject” in any given (clause or predication of a) language. How this is done is a central design feature of any typological study, and the specific operationalizations used in any given study are a matter of “construct validity” (Cronbach & Meehl 1955).

A second important quality criterion of empirical research is reliability. In the case of typological studies, reliability most importantly concerns the consistency of classification: given some property P assigned to some entity x, would different researchers, or even the same researcher at different times, classify x in the same way (with respect to P)? Anyone who has participated in a typological project will be familiar with the sometimes-unexpected difficulties that one encounters when classifying data, even for major variables and well-known languages. Reliability can be tested by running interrater reliability tests, when several researchers annotate the same data set independently, and the degree of (dis)agreement on the coding decisions is determined. While interrater reliability testing is (still) often neglected in linguistic typology, it is regarded as an essential component of studies in psychology and other fields (see for instance Hallgren 2012).

As validity and reliability are primarily a matter of research design, not of research data management, we will not go into any further detail at this point. What matters from the point of view of research data management is transparency in three respects: (i) the comparative concepts have to be defined explicitly, (ii) the operationalizations used have to be precisely described, and (iii) the data analyzed, and the software/scripts used for the analysis, should be made available to the readers (cf. also Holton, Leonard, & Pulsifer, chapter 4, this volume). If conditions (i) and (ii) are met, a study can be said to be reproducible; if, in addition, condition (iii) is met, the study can be said to be replicable (cf. Plesser 2018; cf. also Gawne & Styles, this volume).

Crosslinguistic variation manifests itself in properties of linguistic systems. Typological studies traditionally refer to such systems as “languages,” but there are non-trivial problems with this term. First, what we traditionally call a language is a social construct; from a cognitive point of view, linguistic systems are stored in individual minds, as a part of a language user’s linguistic repertoire. Related to this problem is the question of delimitation between levels of classification such as idiolects, varieties, languages, genealogical groupings, and so on. Moreover, there is no objectivity in describing linguistic data, and any description of any linguistic system implies a fair amount of subjective analytical decisions. Rather than referring to “languages,” typologists have therefore started to use the term doculect for “a linguistic variety as it is documented in a given resource” (Cysouw & Good 2013:342). In practice, typological studies standardly use information from different sources, however, thus effectively subsuming doculects under abstract entities mostly (still) labeled “languages.” As this is not the place to solve the fundamental problem of what entities underlie or constitute the observations in typological research, we will adhere to this established practice in the following. It should be borne in mind, however, that language will be used as meaning “linguistic system as perceived by some (group of) analyst(s).”

Simplifying somewhat, linguistic systems can be regarded as comprising sets of elements (e.g., words) classified into categories (e.g., adjective, feminine, singular), and relations holding between the elements of specific categories (e.g., agreement between an adjective and a noun modified by that adjective; cf. Corbett 2006). In a grammar-based approach, the basic properties underlying crosslinguistic studies are typically generalizations or existential quantifications over elements of the categories constituting the linguistic systems in question (cf. Dryer & Haspelmath 2013 for various examples). For instance, the property of being isolating is a generalization over the form (non-segmentability) of the words of a language (cf. Bickel & Nichols 2013); the property of being a verb-object language is a generalization over the order of constituents of specific categories, and with specific functions, in a language (cf. Dryer 2013); and the property of having a velar nasal is an existential
Managing Data in a Typological Study

quantification over the sound inventory of a language (cf. Anderson 2013). In such studies, properties are thus attributed to languages, and we can make statements such as “English has a velar nasal.”

Typological studies dealing with the realization of functional domains such as the future tense are often based on existential quantifications as well, such as when they determine whether or not a language has an inflectionally realized category expressing the domain in question (cf. Dahl & Velupillai 2013). Such studies often assume a one-to-one relationship between languages and (the existence of) grammatical subsystems expressing the relevant functional domains (“language L has/does not have a future tense”). In other functional domains, such one-to-one relationships cannot be assumed. For example, languages may have various types of impersonals, the phenomenon documented in the ImproType database described in section 3 of this chapter. French uses on as an impersonal pronoun, cf. (1) (see for instance van der Auwera, Gast, & Vanderbiesen 2012; Gast & van der Auwera 2013; Gast 2015 on impersonals).

(1) French on

On ne vit qu'une fois.

IMP NEG lives but one time

“One/you only live(s) once.”

Some languages use verbal strategies to describe actions or events without specifying the participants, cf. the Estonian examples in (2).

(2) Estonian

a. Tullakse ja minnakse.

come.PRES.IMP and go.PRES.IMP

“They (people) come and go.”

b. Siin ehitatakse uut maja.

here build.PRES.IMP new.PART house.PART

“They are building a new house.”

(Blevins 2006:238, referring to Tuldava 1994:372)

Polish often uses the reflexive pronoun się (cf. (3)), but it also has other means, such as the impersonalizing suffix -no, cf. (4).

(3) Polish

Sie

Żyje się tylko raz.

live.3SG REF. only once

“One only lives once.”

(4) -NO

Znowu podnieśli no podatki.

again raise-NO taxes

“They’ve raised taxes again.”

As the Polish examples illustrate, there is no one-to-one relationship between languages and “expression strategies” for a given domain or category. Consequently, properties can only be attributed to expression strategies, not to languages. Note also that typological studies may be interested primarily in correlations between formal and functional properties of expression strategies, or the division of labor between the strategies available in a given language (cf. section 4). While it is relevant that a given linguistic expression strategy $S$ is an element (or combination of elements) of some language $L$, the typologically relevant properties are primarily attributed to $S$ in such cases, not to $L$. The data structure needed for this type of study can thus be illustrated as shown in figure 53.1.

Properties can be stored in the form of attribute-value pairs, represented as “ATTRIBUTE: value” in the following. For example, if the main exponent of an impersonalization strategy is a verbal affix, it can be said to have the property “CATEGORYOFEXponent: verbal.affix.”

3 Managing typological data in databases:
The Typological Database of Impersonals (ImproType)

3.1 The structure of the database

The ImproType database grew out of a project jointly carried out by research groups from the Friedrich Schiller University of Jena and the Université de Paris VIII. This database is an instance of the Extensible Linguistic Database (XLD) system developed by Dimitriadis and van Vugt. Impersonals were introduced with examples from French, Estonian, and Polish in (1)–(4). Given that the notion of “impersonal” covers a broad range of

![Figure 53.1](http://direct.mit.edu/books/book/chapter-pdf/1980037/c047900_9780262366076.pdf)
expression strategies (e.g., pronouns, non-finite forms), it is defined in functional terms as an operation on argument structure, cf. (5).

(5) **IMPERSONALIZATION** is the process of filling an argument position of a predicate with a variable ranging over sets of human participants without establishing a referential link to any entity from the universe of discourse. (Gast & van der Auwera 2013:136)

The definition in (5) can be regarded as a comparative concept. This comparative concept was operationalized using translation equivalence with a set of trigger sentences that were administered in five major European languages (English, French, German, Russian, Spanish). The set of trigger sentences included (inter alia) the examples in (6)–(8), as well as their translations into the other major European languages.

(6) You/one only live(s) once.
(7) They have raised taxes again.
(8) One/you should not drink and drive.

Each language is assumed to have a repertoire of expression strategies, in other words, ways of expressing impersonalization as defined in (5). A database used as a repository for a (grammar-based, functional) typological project of this type minimally thus requires ways of associating expression strategies with a given language and attributing properties to these expression strategies. Moreover, we need to store metadata (cf. also Good, chapter 3, this volume) and examples illustrating the properties attributed to the expression strategies. The basic data model of the ImproType database is shown in figure 53.2, which extends figure 53.1. The arrow between properties (of strategies) and examples indicates that these entities can be linked to each other in the database (cf. section 3.3).

Given that the creation of typological databases has increasingly become a cooperative endeavor, with researchers contributing from all parts of the world, there are a few technical requirements on such databases. Trivially, a typological database should be accessible online. Less trivially, ways have to be found to ensure consistency, in the sense that different contributors should use identical terms for the properties that they assign to observations. One way of ensuring consistency is by restricting the possible values available for some attribute. If a database allows users only to select from predefined values, the database should be extensible, in the sense that new properties (both attributes and values) can be added if need be. As modifications of the database structure should only be carried out by specific users (e.g., from the core project team), a typological database moreover needs appropriate user management facilities. Even though it was mentioned in section 3 that reliability is not primarily a matter of research data management, a typological database can moreover be designed in such a way that it aids enhancing the reliability of the data, for example, by providing instructions and examples.

### 3.2 Properties stored in the ImproType database

The domain of impersonalization has been studied in quite some detail from a theoretical point of view, based on well-known European languages (e.g., van der Auwera, Gast, & Vanderbiesen 2012; Gast & van der Auwera 2013; Gast 2015). Impersonals vary in terms of specific structural and distributional properties. As illustrated in (1)–(4), the expression strategies may belong to different syntactic categories, and they may have different morphological realizations. More fine-grained distinctions can be made, for example, relating to morphosyntactic

![Figure 53.2](http://direct.mit.edu/books/book/chapter-pdf/1980037/c047900_9780262366076.pdf)
properties (such as agreement), and the syntactic functions that can be impersonalized. Moreover, strategies of impersonalization vary in terms of their distribution relative to semantic contexts. For instance, some impersonals (such as English *one*) are mostly found in generic or modal sentences (cf. (9)), whereas others are also used in (non-modal) episodic sentences, for example, German *man* (cf. (10)).

(9) English *one*
   a. Generic/modal
      *One shouldn't drink and drive.*
   b. Episodic
      *One has stolen my bike.*

(10) German *man*
   a. Generic/modal
      *Man sollte nicht betrunken Auto fahren.*
      
   b. Episodic
      *Man hat mein Fahrrad gestohlen.*
      
      “Someone stole my bike.”

The range of distributional properties investigated for each expression strategy cannot be explained in detail here (cf. van der Auwera, Gast, & Vanderbiesen 2012; Gast & van der Auwera 2013). Suffice it to say that three dimensions of classification (i.e., attributes) figure centrally in the distribution of impersonals: (i) event quantification (episodic, habitual, generic), (ii) quantification over participants (existential/singular, existential/plural, collective, universal), and (iii) referential restrictions (exclusive/unrestricted, exclusive/restricted/external, exclusive/restrictive/internal, inclusive). These properties can be cross-classified, yielding a feature grid with $3 \times 4 \times 4 = 48$ cells or nodes as shown in figure 53.3. The gray shades distinguish feature combinations that

![Figure 53.3](http://direct.mit.edu/books/book/chapter-pdf/1980037/c047900_9780262366076.pdf)
are represented in the database (dark gray) from those that are excluded (light gray), either because they are logically impossible or because they are not sufficiently distinct from neighboring nodes (i.e., the distinction cannot be operationalized).

For a selection of the nodes shown in figure 53.3, diagnostic sentences were identified, operationalizing the feature combinations in question. For example, node 48 is associated with the features “generic,” “universal,” and “inclusive” and thus represents sentences of the type One only lives once (6), as this is a generalization over all human beings including both interlocutors.

3.3.1 Add user A user with a user identifier and a password was added by a project member with “account management” rights, mostly one of few administrators. Users providing data will normally have rights for “simple property management” (the other options are “entity management” and “all property management”).

3.3.2 Add answerset Answersets (corresponding to a language or doculect) can be added by users with “entity management” rights. These users will mostly be project members familiar with the structure of the database. The top of the section with information on the answerset is shown in figure 53.4. The answerset in question is here called “poli1260LJ,” using the Glottocode for Polish (poli1260, cf. Hammarström, Forkel, & Haspelmath 2019) with an identifier for the expert (the data were provided by Łukasz Jędrzejowski).

3.3.3 Provide metadata The following metadata was gathered:

- Language: variety, Glottolog code, ISO 639-3 code (cf. Eberhard, Simons, & Fennig 2019), any transliteration conventions, consulted sources
- Type of analysis: single consultant analysis, analyst working with a consultant, consultant-analyst (plus a second analyst), based on written sources only
- Analyst conducting session: name, analyst’s familiarity with the language
- Interview sessions: dates, interview languages
- Consultant: name, date of birth, gender, occupation, level of education, level of proficiency in the analyzed language, language biography, age of acquisition, familiarity with the grammar of the language
- As a point of reference for the impersonals, the most important existential and universal indefinite pronouns were moreover elicited (e.g., Polish ktoś “someone,” wszyscy “everyone”).

### Table 3.3.3.1

<table>
<thead>
<tr>
<th>Id</th>
<th>poli1260LJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Polish (Standard)</td>
</tr>
<tr>
<td>Description</td>
<td>Standard Polish as described by Łukasz Jędrzejowski</td>
</tr>
</tbody>
</table>

### Table 3.3.3.2

<table>
<thead>
<tr>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Language name</td>
</tr>
<tr>
<td>Comments</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. Glottolog code</th>
</tr>
</thead>
<tbody>
<tr>
<td>poli1260</td>
</tr>
<tr>
<td>Comments</td>
</tr>
</tbody>
</table>

**Figure 53.4**

Input screen for properties of the answerset (doculect).
3.3.4 Identify the impersonalization strategies of the language under analysis

To identify the most important impersonalization strategies of the language in question, trigger sentences such as those in (6)–(8) were administered. Moreover, consultants were asked specifically if a form of the second-person singular, or the third-person plural, could be used impersonally. For Polish, six expression strategies for impersonalization were identified in this way. They are illustrated in (3), (4), and (11)–(14).

(11) **TO**
Przebi- to mu opony.
puncture-**TO** him.DAT tires
“They punctured his tires.”

(12) **NO-SIE**
Nie powin- no się kła mać.
NEG should-**NO REFL** lie-INF
“One should not lie.”

(13) **3PL**
Znowu podnie śli podatki.
again **raise.3PL-PTCP.VIR** taxes
“They’ve raised taxes again.”

(14) **2SG**
Bilet kupujesz w autobusie.
ticket buy.**2SG.IND** in bus.LOC
“You/one (can) buy(s) a ticket in the bus.”

Each of the strategies was then added to the answer-set. For each strategy, a typical example is required. The input screen for examples is shown in figure 53.5. Examples are associated with one strategy/marker, and they are classified in terms of grammaticality. The range of values for the variable “grammaticality” can be customized (like any other value for a given attribute) by users with “all property management.” There is also a comments field for prose comments (not shown in figure 53.5).

Examples can (but do not have to) be linked to coding decisions. In terms of the data model, this means that each example will be associated with a set of attributes corresponding to properties of the strategy in question, for example, GenUnivIncl for “generic, universal, inclusive.” Because examples are uniquely linked to expression strategies (such as sie) from answer sets (such as poli1260LJ), such attributes can uniquely identify coding decisions.

The following basic properties were determined for each strategy (mainly properties of the exponent[s]):

- Name and description of the strategy
- Main exponent(s) (e.g., sie)
- Lexical gloss or likely historical source of the exponent (e.g., reflexive pronoun)
- Locus of encoding (e.g., auxiliary, in argument position, on predicate)
- Any specific properties of the verbal predicate (e.g., non-finiteness)
- Obligatoriness of the exponent
- Phonological properties of the exponent (possible stress)
- Possibility of anaphoric reference of the impersonalized argument position

3.3.5 Determine morphosyntactic and distributional properties of the expression strategies

Having described the formal makeup of the expression strategies, we can determine their distributional properties. ImproType divides these properties into two major sections, morphosyntactic properties and semantic properties. The section on morphosyntax requires a complete paradigm of the exponent(s) and contains questions relating to the following:

- Categories of gender, number and case
- Syntactic functions that can be impersonalized
• Binding properties (possessive determiners, reflexives and reciprocals, binding across clauses, binding into a purpose clause)

As has been pointed out, for each coding decision, examples should be provided, using the input screen shown in figure 53.5. For instance, the questionnaire contains the question “Can the impersonalized argument bind possessive determiners in the same clause (e.g., One/you, should not waste one’s/time)?” If this is possible, as in English, the answer will be “yes” (TRUE), and a grammatical example should be provided; otherwise the answer is “no” (FALSE), and an ungrammatical example should be given.

The section on semantic properties is subdivided into questions about universal readings and questions about existential readings. This is only a matter of exposition (configurable in the database), and from the point of view of data structure, all properties are simply attributed to the expression strategy in question. This section is organized around sentences that are defined in terms of the semantic features introduced and summarized in figure 53.3. Figure 53.6 shows the input screen for node 16, which stands for the features “exclusive/restricted/external,” “episodic,” and “universal.”

Figure 53.7 illustrates the structure of the ImproType data, adding some examples to the diagram in figure 53.2.

4 Retrieving and analyzing the data

The database has a search mode, which allows users to search answer sets, strategies, and examples, using all the properties stored in the database as filters. The search form is generated dynamically; that means that if a new attribute is added, or a new value for an attribute, these attributes and values will also be displayed in the search form. Figure 53.8 shows a part of the section for “Impersonal strategies” from the database.

To carry out quantitative studies, the data can be exported (cf. also Han, chapter 6, this volume). While the database itself does not have an export function,
Managing Data in a Typological Study

Once the data have been exported, they can be analyzed using quantitative methods. Given the relatively low coverage of lesser-described languages in the database, robust quantitative statements cannot yet be made. We will therefore restrict ourselves to illustrating the type of question that can be addressed with the data from ImproType in the following.

a Python script for the export of data from any XLD database is available on GitHub. So far the data have been described in an abstract format, in the form of tree diagrams with attribute-value pairs as their leaves. Such attributions of properties can be stored in various ways. The XLD system underlying the ImproType database uses a relational database system (MySQL) for that purpose, where the data are distributed over various tables. As the database is fully extensible—new attributes and values can be defined for the entities “answer set” and “strategy”—these attributes and values are also stored in the database (and recovered dynamically for any operation, either input or output). The export script for XLD databases returns the data in the form of JavaScript Object Notation (JSON) objects (cf. Han, chapter 6, this volume). JSON objects can be regarded as sets of attribute-value pairs. They have the form shown in (15).

(15) \{attribute-1: value-1, attribute-2: value-2 . . . \}

Given that JSON objects may have other JSON objects as their values, a treelike structure underlying the data of the ImproType database can be stored in the JSON format as shown in (16).

\} \\
\} \\

Once the data have been exported, they can be analyzed using quantitative methods. Given the relatively low coverage of lesser-described languages in the database, robust quantitative statements cannot yet be made. We will therefore restrict ourselves to illustrating the type of question that can be addressed with the data from ImproType in the following.
Conclusions

This chapter has intended to provide an overview of the ways in which the data of a typological study can be managed using a relational database system such as ImproType. While we believe that typological databases are extremely useful, in various ways, it should have become clear that many challenges remain, some of them theoretical or methodological, others technical.

We pointed out in section 2 that, in general, empirical research has to meet two important quality criteria, validity and reliability. Finding clear definitions and precise operationalizations is a question of research design, and a database can only be of limited use for these tasks. Note that our way of dealing with the problem of cross-linguistic comparability as discussed in section 2—using translation equivalence with sentences from major languages as operationalizations of the variant properties captured in the database—is certainly not ideal. Obviously, multimodal stimuli would be helpful and are in fact increasingly used in typological studies (cf., for instance, Fedden, Brown, & Corbett 2010).

While validity is essentially a question of research design, not research data management, a database can help researchers to ensure a certain level of consistency and reliability. For example, it can provide explanations and examples, as a part of a protocol (cf. Holton, Leonard, & Pulsifer, chapter 4, this volume) built into the software as it were, and it can restrict the range of values available for a given attribute. Still, the reliability of the coding decisions would have to be checked with interannotator agreement tests, to see to what extent the operationalizations are robust. While the database provides an infrastructure for this—different answer sets could be filled in by various specialists—we have not been able to carry out such tests, partly because financial or time resources were lacking. The problem is of an even more general nature, however: in linguistic typology, data from specific languages can often only be provided by very few specialists, sometimes just a single scholar. In such cases no interannotator reliability testing is possible. The distributed nature of the data collection process, with contributions made by various colleagues from all parts of the world, makes it difficult to gather sufficient amounts of data even for languages with a high number of language users or specialists. One way of dealing with this problem in our project was by having data sets double-checked by...

Figure 53.9
The semantic map proposed by Gast and van der Auwera (2013).

One of the central topics in semantic typology is the mapping from form to function. We assume that linguistic markers are semantically homogeneous in the sense that they cover a set of meanings that are directly related to each other. In the semantic map framework (Haspelmath 1997; van der Auwera & Plungian 1998; Gast & van der Auwera 2013), meanings are assumed to form a network, and markers are hypothesized to cover a contiguous region in that network (the "contiguity requirement"). Gast and van der Auwera (2013) have proposed the semantic map shown in figure 53.9.

The nodes of figure 53.9 correspond to sentences of the following type (the number category of the verb depends on the number category of the relevant impersonal):

1. IMPRS has/have stolen my car.
2. IMPRS has/have surrounded us.
3. IMPRS has/have raised taxes again.
4. IMPRS eat(s) dragonflies in Bali.
5. IMPRS only live(s) once.
6. IMPRS should not drink and drive.
7. What happens if IMPRS drink(s) sour milk?

The data in the ImproType database can be used to test the semantic map shown in figure 53.9. If a language has a strategy that violates the contiguity requirement, the hypothesized semantic map can be regarded as falsified. While no such cases have come to our attention yet, the database still contains relatively few non-European languages, as the process of data gathering is time-consuming and mostly requires native linguists.

5 Conclusions

This chapter has intended to provide an overview of the ways in which the data of a typological study can be managed using a relational database system such as ImproType. While we believe that typological databases are extremely useful, in various ways, it should have become clear that many challenges remain, some of them theoretical or methodological, others technical.

We pointed out in section 2 that, in general, empirical research has to meet two important quality criteria, validity and reliability. Finding clear definitions and precise operationalizations is a question of research design, and a database can only be of limited use for these tasks. Note that our way of dealing with the problem of cross-linguistic comparability as discussed in section 2—using translation equivalence with sentences from major languages as operationalizations of the variant properties captured in the database—is certainly not ideal. Obviously, multimodal stimuli would be helpful and are in fact increasingly used in typological studies (cf., for instance, Fedden, Brown, & Corbett 2010).

While validity is essentially a question of research design, not research data management, a database can help researchers to ensure a certain level of consistency and reliability. For example, it can provide explanations and examples, as a part of a protocol (cf. Holton, Leonard, & Pulsifer, chapter 4, this volume) built into the software as it were, and it can restrict the range of values available for a given attribute. Still, the reliability of the coding decisions would have to be checked with interannotator agreement tests, to see to what extent the operationalizations are robust. While the database provides an infrastructure for this—different answer sets could be filled in by various specialists—we have not been able to carry out such tests, partly because financial or time resources were lacking. The problem is of an even more general nature, however: in linguistic typology, data from specific languages can often only be provided by very few specialists, sometimes just a single scholar. In such cases no interannotator reliability testing is possible. The distributed nature of the data collection process, with contributions made by various colleagues from all parts of the world, makes it difficult to gather sufficient amounts of data even for languages with a high number of language users or specialists. One way of dealing with this problem in our project was by having data sets double-checked by...
Managing Data in a Typological Study

core team members and getting back to the informants in case of doubt with respect to specific coding decisions, but we cannot rule out that the database may contain inconsistencies. The fact that we have not been able to quantify interannotator reliability for the data is a serious drawback when it comes to publishing results based on the database, as interannotator reliability testing is more and more becoming a standard in international publishing. Given the general nature of this problem, formulating standards of empirical research specifically for typological studies would be a welcome step in turning linguistic typology into a full-fledged empirical paradigm (cf. also Gawne & Styles, chapter 2, this volume).

Notes

1. “But what an achievement it would be were we able to confront a language and say to it: ‘you have such and such a specific property and hence, also such and such further properties and such and such an overall character’—were we able, as daring botanists have indeed tried, to construct the entire lime tree from its leaf. If an unborn child could be baptized, I would choose the name typology” (von der Gabelentz 1891:481, translation quoted from Gast 2011:21).

2. “Unlike descriptive categories, [comparative concepts] are not part of particular language systems and are not needed by descriptive linguists or by speakers. They are not psychologically real, and they cannot be right or wrong. They can only be more or less well suited to the task of permitting crosslinguistic comparison” (Haspelmath 2010:664).

3. The Typological Database of Intensifiers and Reflexives is an example of an early database implementing this type of data model; cf. Gast et al. (2007), Gast (2009).

4. In corpus-based approaches, properties of the type described in section 3 are not attributed to languages or strategies, but to markables, i.e., (sequences of) tokens of linguistic output as recorded in a corpus.

5. URL: https://linktype.iaa.uni-jena.de/ImproType. Guest login: “guest”, password: “ImproGuest.”


References


