Managing data in TerraLing, a large-scale cross-linguistic database of morphological, syntactic, and semantic patterns

Hilda Koopman and Cristina Guardiano

[version of 8-17-2020, links corrected from http://test.terraling.com to http://terraling.com]

1. Introduction

TerraLing (http://terraling.com/) is a database-backed web application set up to collect, store and explore data for comparative research in the linguistic sciences. TerraLing is publicly accessible and open-ended: new languages, contributors, properties and databases can be added so as to allow the database to grow over time. Its basic setup allows working with linguists who are native speakers or signers as language-experts providing the data. This gives researchers the opportunity to use the tools of theoretical linguistics to access the implicit knowledge of native speakers/signers to probe the cross-linguistic situation. The basic database schema is flexible, which means it can be adapted to the research needs of individual researchers. TerraLing aims to: (a) make linguistic data widely available on a group of sister databases, whether the data come from well-studied or understudied languages (including dialects), from spoken or signed languages, or from endangered, extinct, or emerging languages; (b) to provide a common set of powerful queries and analytical tools on the web application to explore the data in

each database, and (c) to enable language researchers to easily set up additional sister databases. The long-term goal is to turn TerraLing into a ready-made community tool that linguistic projects can use to gather and store their data for comparative research purposes.

1.1 Brief history of TerraLing

TerraLing is the result of a collaboration of linguists and computer scientists from NYU and UCLA over the past decade. It is currently led by Hilda Koopman (UCLA Linguistics), and is based on original ideas of Chris Collins and Richard Kayne (NYU linguistics), who envisioned a publicly accessible, open ended, language expert-oriented internet database (described in Collins & Kayne 2007). This vision defined the basic functionality of the database (as described in section 2), designed by Dennis Shasha (NYU, Computer Science).

TerraLing was built from scratch. An NSF-funded[1] pilot web application was launched in 2009 around the language-expert SSWL database (Syntactic and Semantic Structures of the World's Languages).[2] Based on the lessons learned from

[1] NSF SGER: Prototype and Specifications for a Web-based Database of the Syntactic Structures of the World's Languages (SSWL). BCS, 0817202, $68,133 with supplement, Chris Collins, PI.

[2] With Semantic added to the original name.

the original prototype, it was reprogrammed as TerraLing, with Marco Liberati and Hannan Butt at the backend.[3] As it failed to secure further NSF funding, the overall project slowed down, but programming continued on a volunteer basis, supervised by Dennis Shasha,[4] and linguistic development continued in the background.

In July 2017, TerraLing was sufficiently developed for migration of the original SSWL database, and for hosting new databases.

Further work on the backend, search tools, user interface and administrative interface is in progress. As of September 9, 2019, TerraLing hosts four databases:

- SSWL
- Conjunction and Disjunction
- Anaphora
- Cinque's Universal 20 Database (see 3.3 for further discussion).[5]

---

[3] With the help of many programmers over the years as listed on the website.

[4] With some financial support of various UCLA Faculty Research Grants and a grant of the Truus and Gerrit van Riemsdijk Foundation, hereby gratefully acknowledged.

[5] SSWL (http://terraling.com/groups/7) is managed by Hilda Koopman, UCLA Linguistics; Conjunction and Disjunction (http://terraling.com/groups/8) is a semantic typology database, led by Viola Schmitt and her project members, University of Vienna, Linguistics; Anaphora (http://terraling.com/groups/9) is led

1.2 Rationale

The idea of an open-ended,[6] language expert-oriented, internet database was borne out of a general need for a tool that can support theoretically guided research.

Formal syntacticians and semanticists consider data about the properties of individual languages. However, it is necessary for further progress of the field to find out what generalizations hold across languages and why they hold, what properties can vary and why, what properties are invariant across languages, what properties correlate, and what gaps there are: this will invariably help narrow down the set of hypotheses that we entertain about the language faculty.[7]

Future theoretical progress in formal syntax/semantics thus demands that we move towards theory-oriented thinking, to make precise empirical predictions of different proposals that can be tested on comparative data for as many languages/dialects possible. This enterprise requires that we make inventories about

---

by Dominique Sportiche, UCLA Linguistics; Cinque's Universal 20 Database (http://terraling.com/groups/15) is a conversion of Guglielmo Cinque's private database (University Ca' Foscari, Linguistics).

[6] Cf. Gawne & Styles, this volume.

[7] See for example Kayne (2013), Cinque (2005) and references cited there.

what is found (at the necessary level of granularity), a task that not only serves theoretical linguistics, but linguistics in general, as well as related fields.

This novel method of investigation crucially requires novel tools and novel ways of collecting the data. To do so successfully, we have to import a basic methodological tool of formal linguistics, which is the use of introspection, that has driven most results in formal syntax and semantics so far.

Work in the description and analysis of the structure and distribution of linguistic diversity across time and space, as well as on the internal structure and nature of human language, has produced a huge amount of empirical/typological data from different types of languages. These data should be made available in some repository form and be accessible to the general public to be useful.

Within the typological and documentary linguistics tradition, new data has been made available through many new reference grammars of individual languages (many previously undescribed), as well as through publicly available databases like the famous online global database WALS (Dryer & Haspelmath 2013), which collects the research results of 55 authors and is based on research spanning many decades. Yet, for all their virtues, descriptive grammars, WALS, or *corpora* for that matter, are not sufficient to answer the questions for theoretically guided research, access to native speakers/signers' intuitions is crucially required as well. To probe the structure of some sentence, we need to apply a battery of diagnostic tests to it. Such tests produce constructed examples, which require

judgments as to whether these are acceptable or not (with a certain meaning, and given a particular environment). These will have to include intuitions on constructed examples that control for a number of variables.

The data coming from the formal linguistic tradition have been made available in some repository form accessible to the general public are mainly: (i) new descriptive grammars (based to a large extent on introspective judgments), and (ii) databases that document microvariation, i.e. variation found in closely related varieties of languages/dialects.[8] The research on languages that are widely and intensely studied continues to yield an astounding amount of new knowledge (and the end is nowhere in sight). This can be measured by the new descriptive grammars that resulted almost exclusively from the generative syntax toolkit,[9] for example the 2000 pages of Huddleston and Pullum (2005) on English, the 5000 pages on Bosque and Demonte's (1999) Spanish grammar, or the eight (open access) volumes of new description of Dutch syntax written by Hans Broekhuis and collaborators (for example, Broekhuis & den Dikken 2012; Broekhuis 2013;

[8] http://www.dialectsyntax.org/wiki/Projects_on_dialect_syntax.

[9] Acceptability judgements in fact do require detailed contextual information (cf. Good, this volume): examples should always be considered in context (see the Coordination and Disjunction Database for detailed examples: http://terraling.com/groups/8).

Broekhuis & Corver 2016). Furthermore, quite remarkably, native speakers seem to agree on the vast majority of these data. As we stress, these descriptions mostly result from introspective data guided by the ever larger number of diagnostic tools and methods that formal linguistic theory provides[10] (and of course, these descriptions build on previous grammars, general literature, corpora, any data that exists, as well). This raises the question how reliable introspective data are. Though introspective methods are often considered to be unreliable, it is important to point out that this is not confirmed by experimental research. Quite the contrary, Sprouse, Schütze, and Almeida (2013), Sprouse and Almeida (2012), and Schütze and Sprouse (2014) have experimentally tested the data in journal publications or textbooks discovered by these methods, and proven they are highly replicable (between 95% and 98% depending on the datasets). Thus, to enable theoretical progress on the basis of cross-linguistic data, we should be able to access the intuitions of native speakers or signers: TerraLing allows doing so (though nothing forces to gather data in this way).

Against this background, we now turn to section 2, which discusses the database functionality and describes the flexible database schema, the search interface, and the basic management set up. Section 3 is written as a guide for the

[10] See in particular Sportiche, Koopman, and Stabler (2013, Chapters 3, 6, 7, 11, 12).

reader who would want to develop content for one of the existing databases, set up their own database within TerraLing for a cross-linguistic project, or who would otherwise want to be involved in the general project. Section 4 discusses data collection, and various issues related to glosses, and academic credit. Section 5 provides a short summary of the chapter.

2. Basic database functionality and description

To meets its goals, TerraLing and its database(s) must meet the following functionality:

1. They must persist and expand over time, and be openly accessible worldwide.
2. They must allow flexible additions to data as new properties and new languages are added, without any need for reprogramming.
3. The web application must allow disciplined and secure curation of data by multiple linguists.
4. The data stored in the database(s) must be easily extractable and usable for exploration and research purposes.

We briefly discuss below how this is achieved, and present a basic description of the (flexible) database schema.

2.1 Durability and accessibility

TerraLing and its databases must be able to last over time, and be openly accessible worldwide.[11] TerraLing is currently hosted on the highly secure industrial site ACS (Amazon Cloud Services), and is accessible worldwide. Regular automated back-ups further protect the data.[12] While the default option is to be openly accessible, it is also possible to restrict access to a database, if so required. Each individual database has a toggle for a *private* or *public* setting. A private setting restricts access to a group of researchers for a specific duration (for example, for the duration of a funded project). A simple switch to a public setting will make the database publicly accessible.

2.2 Basic design

The project builds on a simple but flexible property-as-value model, which professor Dennis Shasha has used successfully in his work in plant genomics.[13]

---

[11] See Kung, this volume.

[12] See Han, this volume.

[13] TerraLing is built on Ruby on Rails, one of several web development frameworks that support database-backed cross-browser web applications and enjoy strong open source community support. Because Ruby on Rails embodies a model-view-controller paradigm, changes can be quickly deployed on a browser,

This model ensures that new content can be added over time. Linguistic data are characterized by data linked to *objects of description* (typically languages or dialects). Each such object can be characterized by a set of *property-value pairs*.[14]

The information stored in the database is represented through four types of tables:

1. Languages (*languagename*, *propertyname*, *value*, *contributorname*, *date*, *time*)
2. Properties (*propertyname*, *description*, *contributorname*, *date*, *time*)

---

first on the programmer's laptop and then on the web using Capistrano. The model-view controller design pattern allows different sites to share the same data model (same database schema) but different user-visible names (different views). Ruby on Rails and the backend database we have chosen MySQL are open source and free, thus lowering the barriers to entry. The search interface is implemented with a javascript api that queries the existing rails service. The database software is freely available on GitHub.

[14] In our system, all properties can be reduced to binary values. This will be shown in more details in section 3, where we also illustrate the effects of this choice in terms of accuracy of the typological variation which underlies any analysis.

3. Examples (*languagename*, *sentenceid*, *type*, *propertyname*, *value*, *comment*, *contributorname*, *date*, *time*)
4. Contributors (*contributorname*, *affiliation*, *username*, *password*, *e-mail*, *date*, *time*)

The *Languages* table gives the values for each property. For example, there is a SSWL property for *Predicative adjective agreement* (*Pred_Adj_Agr*).[15] The value of this property can be *Yes*, *No*, NA (Not Applicable). For French and Icelandic the value is *Yes*, and for Dutch it is *No*. A complete listing of properties with their values and accompanying examples for a language is equivalent to a rough grammatical sketch for the areas the properties cover.

In the *Properties* table, each property is associated to a description. *Property Descriptions* have a specific format (as discussed in detail in section 3.4) which provides the definitions of the values of the property and presents a concrete example of a property development.

Evidence for a property setting is given in the form of examples, stored in the *Examples* table. Each example consists of a line of text, a gloss (we recommend using Leipzig glossing conventions whenever possible, but see 4.2 for further comments), a translation, and a comment field. The comment field allows

[15] http://terraling.com/groups/7/properties/407.

contributors to provide further information, which can include further information about the distribution, or the source of the information for example.

The *Contributors* table contains information about who contributes data to the database (where data means the *Property Description*, property-value pairs, or examples). Users interact with a web interface to add or explore data.[16]

Language:	Basaá[17]			
Example:	malaŋ	má	yé	ma-kéŋí
Gloss:	6.onions	6.SM	BE.PRES	beautiful.F.SG
Translation:	The onions are beautiful			
Comment:	*Here is where a comment would go, or a reference to a source of the data*			
Contributor:	Paul Roger Bassong			(SSWL: sentence_1480)

[16] The interested reader can find further information on how to navigate TerraLing here: https://linguistics.ucla.edu/wp-content/uploads/2017/04/Navigating-Terraling-1.pdf.

[17] Basaá is spoken in Cameroon. SSWL registers the iso- or glottolog code for each language, as well as geo coordinates for mapping purposes. SSWL does not record genetic affiliations (Bantu, A40), which is based primarily on lexical relatedness, while Cinque's Universal 20 database does.

2.3 Experts

TerraLing allows for a completely new take on typological research: as opposed to all other linguistic databases, it works with native linguists who are native speakers or signers (or have a deep knowledge of a language) as language experts. This means that it lets us access the implicit knowledge of native speakers and signers directly on a broad typological scale.

TerraLing is set up so as to enable linguists who are native speakers or signers to sign up as an expert contributor for their language. They may do so individually, or as a group.[18] Experts must be approved before they can provide data, i.e. before they are allowed to set property values and provide examples that illustrate the values. Data are tagged by the name of the expert contributor, and remain under their sole control: experts (but no one else, except an administrator) can change

[18] As we link languages to individual speakers/signers (and to locations, via geo-coordinates), we expect that data provided by a contributor might not correspond exactly to those provided by a different contributors (of the "same" language). If two contributors disagree in assigning property values (and there is no misunderstanding in assigning values), two variants of the language, representing the two contributors' judgements, can be added, in consultation with the administrators. If disagreement only concerns one or few properties, then a comment is sufficient to describe variation.

values, examples, or comments. Experts do not have the power to delete their language.[19]

To further ensure the quality and reliability of the data, experts are sometimes paired up with a "mentor" who provides them with explanations about the *Property Descriptions*, checks the property values and the examples, and provides feedback. We would like to generalize this system in the future as it allows an organized check on the data, in addition to familiarizing the community with the database. Further ways to control data reliability are discussed in section 3.4.4.

To allow for disciplined and secure curation of data by multiple linguists, TerraLing has a role module, which defines the following roles: *Administrator* (site administrator, group administrator), *Language Expert*, *Property Author* and *Member*. Administrators control access to the site or group module. Site and group administrators can assign roles to members, and make members into experts for a specific language for example, or demote experts into members. They have full

[19] A language or dialect in SSWL can therefore be defined as the set of forms and property values that characterize the grammar of a specific individual: the contributor. This notion is similar but more restricted that the notion of doculect (see Gast, this volume). Doculect can be used to refer to a specific set of corpora for example, or an analysis "this language/dialect is classified as VSO by linguist1, but as VOS by linguist2".

control over the site or their group's database, except for the ability to delete the group. Access levels thus depend on a specific role assigned to contributors.

2.4 Usability for research purposes

To ensure maximal usability, TerraLing has built-in search functionality, implemented by a javascript api that queries the existing rails service. The search interface can be accessed from the masthead, and consists of an *Advanced Search* page and a *History* page, where saved searches can be stored.

The *Advanced Search* page allows thousands of simple and complex queries, including universal implications and similarity trees. Any field is searchable, and searches can be combined. Searches can be constrained by (all, or any subset of) *languages*, by (all or any subset of) *properties*, by specific combinations of *values*, etc. Up to six properties can be crossed so as to extract all the relevant data patterns in the database. *Compare* allows comparing up to eight languages for all properties that are entered.

All data or search results can be downloaded in csv format from the *Advanced Search* page, or saved on the application, to be later accessed or rerun. Examples can be downloaded from the *Languages* page in json format.

At the time of writing, the search functionality is being improved and further developed, depending on available means and opportunities. This holds as well as for the user interface, *Property Descriptions*, and how-to document videos. The

database is slowly but constantly being developed: the entry page of each database provides a snapshot of the overall data in that database. Since data entry is continuing, this snapshot changes with time. The total percentage of properties set for each language can be found on the *Languages* page; the number of languages set for each property can be found on the *Properties* page. Our ultimate goal is to code up over time as many languages or dialects as possible from all continents (thus, the family skew is irrelevant). As far as the examples are concerned, it is difficult to calculate the exact number of missing ones: as a matter of fact, not all properties need to be exemplified, because one single example can serve to exemplify many properties.

3. Managing data in the databases of TerraLing

This section addresses readers who may want to develop their own (hypothesis-driven) comparative research project, and use TerraLing to do so. We start out with a general overview of the workflow in section 3.1. In section 3.2, we discuss the details of the different aspects of the development.

3.1 General workflow

The TerraLing database is different from existing databases in the following ways.

(i) Requests for proposals to set up a new database, or propose new properties to an existing database, can be submitted at any time to the board for review. The

board asks experts to review and, if necessary, help to improve the submission. If approved, the new properties are fed into the system and data collection can start. The set of linguistic questions for which relevant data can be collected is thus unrestricted.

(ii) New language experts and new languages (including extinct languages) can be added at any time.

(iii) The search engine allows for simple and complex searches (properties of languages, which can be combined and constrained as necessary) and correlated searches (property or language correlations), but also for more complex tasks like searching for implications and typological gaps. It also includes visualization tools such as maps and similarity trees (see section 2.4). Hence, the system is fine-tuned for cross-linguistic research on theoretical questions.

To insure interoperability between the databases on TerraLing, properties should (preferably) have a particular format (see section 3.4.3). Possible answers are restricted to *Yes*, *No* and NA (Not Applicable).[20] Phenomena under

[20] NA means that the language provides no insight into a phenomenon because it lacks a certain property (i.e., if a language has no subject-verb agreement, any property that follow up on subject-verb agreement are irrelevant). The system also contains blanks, namely no answer is given to a certain property: this usually

consideration must be properly described and illustrated by examples that support the property value to insure data reliability. Each *Property Description* has to specify precise criteria for a *Yes* and *No* answer, without technical jargon (and perhaps provide an example of how to set the two values), so experts from different backgrounds can understand how to apply them to their language. Notions that are often required in properties, such as *neutral context*, are pre-specified in the system and connected via links to the actual queries via a *Glossary*. All these requirements must be met so as to make the task doable, and to generate comparable data (on this topic, see also section 4, and Gast, this volume).

Much of the empirical work in any project consists in formulating research questions and plausible properties that fit the requirements of the database, as we see in the next section 3.2.

3.2 Development of a project

With the description of the database and the general workflow as a background, this section will guide readers who may want to develop their own (hypothesis-driven) comparative research project, contribute to an existing one, and use TerraLing to do so. We start with some examples of existing projects in section 3.3 and discuss

---

happens when a contributor has not yet answered a property or, in the case of ancient languages, when the corpus used does not contain the relevant data.

the steps in the development of a new project up to the collection stage with some real examples in section 3.4.

The resulting project, whether in syntax, semantics or morphology or their interfaces, could aim to:

(i) add further content to one of the existing databases, for example, to *SSWL*, *Conjunction and Disjunction* or *Anaphora*.

(ii) convert some existing dataset into a TerraLing database allowing it to grow further (as for example *Cinque's Universal 20 database*).

(iii) set up a new database on TerraLing to meet the specific goals of a particular research project.

3.3 Examples of hypotheses-driven research currently in TerraLing

We start with a small sample of the various theoretically inspired research projects that can be found in the databases on TerraLing. These projects are at the stage where data is being collected, stored, and explored.

The *Conjunction and Disjunction* database (http://terraling.com/groups/8) explores the semantics of conjunction and disjunction by investigating the cross-linguistic realization of such elements. The research is guided by theoretical hypotheses in semantics concerning the meaning of these elements, and explores

how these hypotheses can be tested on typological data. The project already generated important results.[21]

Within the general large-scale *SSWL* database, we mention the following two theoretically inspired projects:

(1) Anders Holmberg and Craig Sailor explore the syntax of *yes* and *no* and gather data on SSWL to determine how affirmative and negative questions can be answered. Different types of elliptical answers are collected. This is the first-time systematic investigation on answers.[22]

(2) Cristina Guardiano and Hilda Koopman are engaged in a systematic documentation project of the determiner region of noun phrases. The properties are organized around the following variables:

(i) an indefinite, definite (or generic) reading of (unmodified) noun phrases, depending on whether the noun is:

(ii.a) a mass noun, a singular or plural count noun,

[21] See https://www.univie.ac.at/konjunktion/texts.html for results and further information.

[22] See the 22 Property Descriptions starting from Q01_Initial polar Q-marker (http://terraling.com/groups/7properties/445) and Holmberg (2015) which explores the results.

(ii. b) a noun with (intrinsically) unique reference,

(ii.c) a proper name, or a proper name modified by an adjective;

(iii) what syntactic position the noun phrase occupies (object, subject);

(iv) how determiners (when present) are ordered with respect to the noun;

(v) whether the noun is a vocative.

This is an area where we find much cross-linguistic variation, both synchronic and diachronic, with formal properties touching on bare nouns versus determined nouns, and issues related to case, adpositions, demonstratives, classifiers, noun classes, quantifiers, and numerals. There is no other database that systematically records this variation for comparative purposes. This project is ongoing (current data for between 55 and 97 languages depending on the property).[23]

Cinque's Universal 20 database[24] is a conversion into TerraLing format of Guglielmo Cinque's private database (in Word format), of Greenberg's Universal

[23] At the time of writing, there are 70 properties, starting (for object properties) with O 01_Indef Mass_1_Can be bare (http://terraling.com/groups/7/properties/467).

[24] http://terraling.com/groups/15.

20 (Greenberg 1963). Universal 20 concerns the attested cross-linguistic word order patterns of *Dem(onstrative) Num(eral) Adj(ective) N(oun)* in the world's languages. In his influential article, Cinque (2005) tallies the patterns that are attested and unattested cross-linguistically. These turn out to be partially different from Greenberg's original universal: only 14 out of the 4!=24 possible patterns appear to occur. The reason why this is so, Cinque proposes, should be found in the Faculty of Language: the unattested patterns cannot emerge from the rules of the grammar. Since the absence of such patterns is crucial, it is imperative to continue to gather all available comparative evidence, and explore potential counterexamples. This database compiles the available information from many heterogeneous sources (previous databases from various sources, data from grammars, articles, native speaker linguists) supplemented by Cinque's own continued research since his article. All data sources are indicated. There are currently data from 1687 languages in this TerraLing database.

3.4 Getting to the collection stage: From research questions to a table of variation and property descriptions

In this section we guide readers who may want to develop their own hypothesis-driven comparative research project through the various different stages and aspects of content development of a new project, with specific focus on how to get from a research question to the collection stage.

Since the novelty of the database is to explore theory-driven questions with comparative data, we focus here on the development of a specific research question, and the lessons we have learned from setting up such projects. The main challenges lie in developing research questions and translate these into queries that can generate comparable linguistic data in a reliable fashion to seed the database, and – ideally – that allow testing theoretical hypotheses. The database is well-suited for the development of a micro-parametric project, as variation found in closely related languages/dialects provides an important window in the principles of the Language Faculty. Given the fact that a great many properties are shared by closely related varieties, these conditions may approximate those of a controlled experiment.

Development of a project is best done through collaborations or in a seminar or workshop-like environment.[25]

The concrete project we build below concerns aspects of the distribution and interpretation of adnominal adjectives. It builds on the substantial body of knowledge accumulated over the years in the general typological and formal linguistic literature (in particular Dixon 1982; Dixon & Aikhenvald 2004; and

[25] We welcome researchers who would like to get involved in helping to push the many projects in advanced stages of development to the collection stage and become one of the Property Authors. Interested readers are encouraged to get in touch with the TerraLing board at linguisticexplorer@gmail.com.

Cinque 2010, and the references cited therein). A partial questionnaire[26] develops this domain in a TerraLing format.

At the most general level, linguists start with the quite general research question *how to define an adjective cross-linguistically*?, and given a definition, ask the following question:

Q1: *How is an adjective ordered w.r.t the noun*?

Leading to questions why, and what the comparative picture can tell us about Universal Grammar.

This lead to a multitude of sub-questions, for example, the small subsample below:

Q2: *Are adjectives always ordered on one side of the noun*?

Q3: *If a language allows stacking adjectives, is there a universal order of different adjectival classes, and if so, how do we explore this question with comparative linguistic data*?

Developing answers to these questions will show how to set up a table of variation, which in turn will lead to the formulations of *Property Descriptions*. These must be formulated in such a way as to generate reliable comparative data.

[26] https://linguistics.ucla.edu/wp-content/uploads/2017/04/Adjectival-Questionnaire.pdf.

3.4.1 Development of a table of variation and corresponding properties

The development of a set of properties is based on a table of variation, which must capture all relevant differences between languages in the specific domain of inquiry. These have binary values, encoded as *Yes*/*No*.[27]

A single binary property can define at most two types of languages. Adding a second property yields four types of languages (see Table 1), adding a third yields eight types of languages (Table 3), etc. Thus, n properties yield a typology of 2^n potentially different languages.

Table 1: Abstract typology, 2 properties

	PROPERTY 1	PROPERTY 2	LANGUAGE?
I.	Yes	No	
II.	No	Yes	
III.	Yes	Yes	
IV.	No	No	

[27] In this section, we leave NA (Not Applicable) out of consideration.

Table 2: Abstract Typology, 3 properties

	PROPERTY 1	PROPERTY 2	PROPERTY 3	LANGUAGE?
I.	Yes	No	Yes	
II.	Yes	No	No	
III.	No	Yes	Yes	
IV.	No	Yes	No	
V.	Yes	Yes	Yes	
VI.	Yes	Yes	No	
VII.	No	No	Yes	
VIII.	No	No	No	

Data collection is based on the notion of *property*. A property can be described as the smallest visible phenomenon able to capture cross-linguistic structural diversity. Properties are conceived as available, in principle, in any language; thus, they must be defined in theory-neutral terms, i.e. avoiding notions (and related terminology) too strictly connected to a specific theoretical vision/background. Properties are conceived as the empirical manifestations of precise structural phenomena. They must be able, in principle, to represent all possible aspects of diversity manifested by a given phenomenon, and at the same time make any language comparable with any other. Thus, they combine requirements of descriptive cross-linguistic adequacy with the need of in-depth explanation of the structure of individual grammars. One lesson we have learned from developing properties so far is that "binning", i.e. collapsing different properties, must be avoided (as much as possible). Decomposition into ever finer smaller (sub)properties is necessary both to ensure generating comparative data and

to allow their theoretical exploration. In order to attain typological exhaustiveness, as many properties must be formulated as needed to capture the observed space of variation. This is a challenging task.

3.4.2 A concrete example of a coding schema

We can illustrate this procedure with a concrete example, namely the word order properties for A(djective) N(oun) orders, which allow comparing the way this is done in SSWL with WALS (Dryer 2013a). This will serve to make the following points: (i) known variation must come out as a result of the coding (hiding known and easily observable variation is not acceptable), and (ii) coding must be based on easily observable criteria so as to ensure reliability and feasibility. We also demonstrate how to use the TerraLing data schema to capture further variation, in effect developing part of the project mentioned in section 3.4.1.

On the most general level, adjectives can either follow the noun, precede the noun, or do either. In WALS, this translates into one feature, *Order of Adj N*, that has four values: *N Adj*, *Adj N*, *no dominant order* and *only internally headed relative clause*. In SSWL, there are two separate (independent) properties, *Adj N* and *N Adj*, each with two possible values (*Yes*/*No*).

Comparison between WALS and SSWL yields different results. For instance, as shown in Table 3 below, French is classified as *N Adj* in WALS but as *Adj N: Yes* and *N Adj: Yes* in SSWL.[28]

Table 3: A comparison of Adj N orders in WALS and SSWL

WALS		SSWL		
Order of Adj N	*Language?*	*Adj N*	*N Adj*	*Language?*
Adj N	… Bengali …	Yes	No	59 … Bengali …
N Adj	… *French*, Swahili, …	No	Yes	98 … Swahili …
No dominant order	… Tagalog …	Yes	Yes	86 …*French*, Tagalog
Internally headed RC	…	No	No	3 … No Adj

Since French clearly has prenominal and postnominal adjectives (*une jolie petite fleur rouge*, lit: 'a nice little flower red'), the *N Adj* value in WALS is very surprising. This is because WALS utilizes the notion *dominant order*:[29] in French

[28] Count in the SSWL table refers to the number of languages with these values setting at the time of writing.

[29] In Matthew Dryers's online supplement on WALS (Dryer 2013b) a number of statements are given as to how dominant order is determined for the word order properties and why it is adopted. It should be clear from these statements that these are not scientific criteria, nor were they meant to be. The goal is clear: to find some measure (however crude it may be) that allows comparing languages,

N Adj is considered dominant since many more types of adjectives follow the noun than can precede it. This now raises various problems. The first problem with this classification is that in WALS French and Swahili are in the same set of languages. But that is incorrect, since all adjectives follow the noun in Swahili, but not all adjectives follow the noun in French. Thus, the first fault of the dominant order criterion is that it fails to capture typological diversity. The second problem is that it is impossible to give instructions so as to get reliable comparative data. Assume for example we have a hypothetical un(der)described language which is just like French. How would one code such a language? Moreover, it prevents exploring further questions, for instance, about possible regularities of which classes of adjectives precede the noun, and why. Finally, the notion of dominant order reveals (and corroborates) the hidden assumption that a language should be uniform in terms of word order (e.g. all modifiers should precede, or follow, the noun). This assumption is in fact not warranted, since languages are quite generally mixed.

---

regardless of the extent or quality of their documentation. Notice that this type of ambiguities/inaccuracies in coding the data might produce unforeseen consequences when data are used for broader purposes, for instance to infer phylogenetic hypotheses about language evolution that are immediately and widely discussed in the general press.

Coding the variation is therefore important for gaining further understanding in principles underlying linear orders.

The structure of the two relevant properties in SSWL does not need to assume any notion of dominant order: all logical possibilities are represented. In fact, all the four language types generated from the two properties are currently attested in the database (type IV instantiates a language with no adjectives). Thus, Swahili and French come out as different, as desired. However, French and Tagalog come out as belonging to the same type, namely to the set of languages that allow both *Adj N* and *N Adj* order. This is in fact all the information that these two basic properties in SSWL can provide (though additional information can be found in comments). In order to further explore whether there is variation between French and Tagalog (or for that matter for the other languages with these property values), and how it is manifested, further finer-grained properties must be formulated, for instance about the relative order of different classes of adjectives (say *color* or *size* adjectives) with respect to the noun. This is plausible given widespread agreement in the literature that the different classes of adjectives line up according to a universal hierarchy (*subj. comment* > *size* > *age* > *shape* > *color* > *gender* > *nationality* > *material*; see for example Sproat & Shih 1991; Cinque 1994; Dixon & Aikhenvald 2004).

Suppose we add two pairs of new properties in the database: (1) a. Adj_{color} *N*: *Yes/No*, b. *N* Adj_{color}*: Yes/No*, (2) a. Adj_{size} *N: Yes/No*, b. *N* Adj_{size}*: Yes/No*. This allows to capture (i) the fact that French *color* adjectives must always follow the

noun: *une fleur rouge* (lit. 'a flower red') leads to the values Adj_{color} *N*: *No, N* Adj_{color}*: Yes*, and (ii) the fact that basic *size* adjectives like *petit, grand* ('small', 'big') precede the noun in French with the values Adj_{size} *N: Yes*, *N* Adj_{size}*: No*. Tagalog on the other hand allows both orders for *color* and *size* adjectives (Schachter & Otanes 1983), which leads to the values Adj_{color} *N*: *Yes, N* Adj_{color}*: Yes*.

This yields Table 4, where the differences between Tagalog and French come out correctly, as well as a difference between French and Italian with respect to the order of the basic *size* adjectives.[30]

Table 4: A more fine-grained SSWL typology

	Adj_{size} *N*	*N* Adj_{size}	Adj_{color} *N*	*N* Adj_{color}	*Language?*
I.	No	Yes	No	Yes	Swahili
II.	Yes	No	No	Yes	French
III.	Yes	Yes	No	Yes	Italian
IV.	Yes	Yes	Yes	Yes	Tagalog
V.	Yes	No	No	No	Lacks color A
VI.	No	No	No	Yes	Lacks size A
VII.	…				

In principle, the additional two properties generate 16 possible combination of values: a combination of four *No* defines a language with no (*size* and *color*)

[30] Finer distinctions between petit, grand and gigantesque ('gigantic') can be built in by further refining the properties.

adjectives, the combination labeled as V. in the table defines a language without *color* adjectives, but with prenominal *size* adjectives, etc. Which patterns are not attested cross-linguistically will fall out from this data-schema.

A fine-grained collection of data like the one we propose reflects the comparative landscape, and can support theoretical explorations and predictions through the sophisticated search interface.

3.4.3 *Property Authors* and *Property Descriptions*

Once a table of variation, or a hypothesis about the data that allow supporting or refuting a theoretical hypothesis, has been developed, the next step is how to translate these into *Property Descriptions*.

Property Descriptions are formulated by *Property Authors*, and are submitted to the editors. *Property Descriptions* (or queries) define and explain the property, provide restrictions about what (not) to consider, define an elicitation task with clear contexts and scenarios which serve to generate the examples on which the property values are assigned. To help the contributor, *Property Descriptions* must show how the property will be set for English, and present examples of languages that represent the combination of different values.

3.4.4 Feasibility testing

Once an initial set of properties has been developed, the properties are sent out to a group of contributor volunteers that test them on feasibility and provide feedback. Can the task be done on their language? Are definitions clear? If they are unclear or ambiguous can the definitions be improved? Depending on feedback, *Property Descriptions* are further refined or adjusted.

4. Data Collection

After approval, the new properties are pushed on the database, and the collection stage can start. Contributors read the *Property Descriptions* on the *Edit Language* page, produce examples on the basis of the elicitation tasks, and determine the value. In order to save the property value, contributors must indicate their level of confidence in the setting of the value. There are three levels of confidence: (i) *certain* (many properties are completely uncontroversial), (ii) *revisit* (some cases are more questionable, require more thought either because the language does not provide an easy answer, or the property may be ill-defined or not refined enough), and (iii) *need help* which will send a message to the *Administrator* and *Property Author*.

We are currently developing an off-line guided questionnaire format to help streamline the task for the contributors. It really requires two separate skills: an *elicitation* task and a *classification* task (applying criteria to assign values).

Values are illustrated with glossed examples that illustrate the set values with examples from individual languages, and give further information, when relevant.

4.1 Data reliability

There are multiple sources for possible errors in the database. So far, there is no central control mechanism: the data are controlled by hand (by the *Administrators*). New data go in a queue to be checked. Reliability of the data is ensured through the adoption of the following measures.

Language experts must be approved by the *Administrators*, and each property value and example are tagged by contributor: therefore, it is always possible, when checking the data and values provided, to interact with the contributor, ask for explanations, further examples, corrections etc. Errors can be corrected at any point, and comments can be added to explain specific value settings.

Each *Property Description* must also be approved by the *Administrators*, in order to make sure that they conform the general guidelines. The main strategy is to lower the chances different types of mistakes can be made at the entry level. Measures include (at present): (i) making the task easy and small (breaking down questions in small parts); (ii) avoiding "binning" (causes cognitive overload); (iii) giving clear instructions, and illustrations in the form of examples, etc.; (iv) presenting all relevant information on one page to minimize the chance shortcuts or guesses are made; (v) making a contributor reflect on their confidence in the

value (contributors must indicate their level of confidence to save the value). Avoiding technical jargon is also crucial in order to make the Property Descriptions accessible to contributors. Confusion happens in particular when standard terminology (like *case*, *agreement*, *clitics*, *bare nouns*, etc.) is used and not defined. As a matter of fact, these terms cover different phenomena in different linguistic communities and traditions, and contributors will be biased according to the uses they have in their respective communities. Consequently, it is crucial that all technical terms are defined with no ambiguity. Obviously, contributors must read and use these definitions of technical terms, and not take shortcuts (i.e. make sure that they understand the meaning of the terms used in the *Property Description*. Overall, it becomes clear quite quickly (from general low confidence scores, and problems around terminology, or low number of answers) which properties are prone to present problems and need extra attention. Another very useful guide for contributors comes from the *Examples* which illustrate each property value: properties not illustrated with examples are potentially problematic; properties usually not problematic are those which can be easily verified because they are part of the general knowledge base.

Since for the contributors of underdescribed or low-density languages the task is inherently more difficult (there are less possibilities for independent control) we have the "mentor" system mentioned in section 2.3.

Once data are entered, the strategy is to make corrections easy: it is frequent for contributors to auto-correct their values or examples. Properties that are answered *Yes* are easy to judge: it is in general sufficient to present a (productive) example to earn a *Yes* value. Properties that have a *No* value are more problematic. Further strategies include: (i) checking examples to see if they illustrate the value; (ii) having a feedback system to identify possible errors and weed them out (originally, SSWL had a forum feature which was set up for that purpose, but since it got little use, we have not reprogrammed it in TerraLing); (iii) enlisting the community (property contributors, administrators, mentors, local community with a common research or areal interest, etc.) to explore the data: this invariably brings up questions, and errors; (iv) finally, we find that search functions are very useful to identify potential outliers, which could turn out to be mistakes or reflect genuine differences. For example, the search function *Compare* is a useful tool to identify potential outliers.

4.2 Examples and glossing

Our guidelines are that for languages with written orthographies, examples are entered into the standard orthography of the language; if there is no standard orthography, examples are entered in the orthography that has been adopted in the linguistic community for the language (or related languages).

As for glossing, our guidelines recommend using the Leipzig glossing conventions[31] but we have not systematically enforced this. This is in part because there are problems with the glossing and naming conventions, where linguists have a strong tendency to (mis)take glosses for analyses. Furthermore, different local communities have developed their own glossing dialects and descriptive terminology.

4.3 Citation guidelines

In this section we provide some practical information about citation guidelines for academic credit, CV, and personal statement.[32] A cite key for the main group pages is currently in development.

Our citation recommendations are listed below.

1. The general work

Koopman, Hilda. Dennis Shasha, Hannan Butt and Shailesh Vasandani (2017 -), TerraLing, http://terraling.com, Accessed on [DATE].

[31] https://www.eva.mpg.de/lingua/resources/glossing-rules.

[32] See Conzett & De Smedt, this volume.

2. Each individual database:

Please follow the information found on the group page. The order of names is left open to the administrators of each database.

[TEAM LEADER/EDITORS],[33] [DATE STARTED -], [NAME OF DATABASE], [URL], Accessed on [DATE].

Examples:

SSWL

Koopman, Hilda (ed.) (2012 -), SSWL, The Syntactic and Semantic Structures of the World's Languages, http://terraling.com/groups/7, Accessed on [DATE]

Conjunction and Disjunction:

Schmitt, Viola, Enrico Flor, Nina Haslinger, Eva Rosina, and Magdalena Roskowski (2017 -), Conjunction and Disjunction, http://terraling.com/groups/8, Accessed on [DATE].

[33] We leave it open to the managers of each database if this is the name of the project leader/editor, the names of the team, or community, or "et al".

3. Property Authors

A considerable amount of research goes into the development of *Property Descriptions,* queries, glossary entries: *Property Authors* must be cited when you use their definitions and schemas of variation.

Examples:

Andrea Cattaneo, Chris Collins, Jim Wood (2011), Property Description for Predicative Agreement, in *SSWL*, https://terraling.com/groups/7/properties/407. Accessed on [DATE].

Viola Schmitt et al. (2017), Glossary entries for Coordination and Coordination, https://github.com/terraling-glossary/glossary/wiki/Coordination. Accessed on [DATE]

Cristina Guardiano and Hilda Koopman (2015 -), A group of 70 properties about properties for the Determiner region, in SSWL (http://terraling.com/groups/7). Accessed on [DATE]

4. Language Experts and Examples:

We recommend that if you use datasets in presentations or written work, you acknowledge the language experts who contributed the data in the datasets in footnotes.

For citing examples we propose the following schema:

Language: Basaá

Example: malaŋ má yé ma-kɛ́ŋí

Gloss: 6.onions 6.SM BE.PRES beautiful.F.SG

Translation: The onions are beautiful

Comment: *Here is where a comment would go, or a reference to a source of the data*

Contributor: Paul Roger Bassong (TerraLing SSWL: sentence_1480)

For in-text references, we propose something like the following: 'As shown by Paul Roger Bassong for Basaá (Terraling, SSWL: sentence_1480), ...'

In-text citations for language data can be as elaborate as the author feels is necessary to make a point, but we encourage generous and inclusive citations, and sending contributors a note to this effect.

If language data ultimately comes from some source other than SSWL (e.g., theses, published paper, monograph, website, etc.), then that source should be cited as well.

4.4 CV, webpage, and research statement

Contributors should record the details of their contributions on their CV, Webpage/Project Webpage, and Research Statement.[34] For *Property Authors*, *Language Experts* and *Administrators*, we recommend the following:

Property Authors

Property Authors should put links to *Property Descriptions* and glossary entries as illustrated in section 4.3 under the heading *Web publications* on their CV, personal webpage, and project webpage. Submit these in their dossiers for the purposes of hiring, promotions, and refer to these in their research statements stating the nature of the work involved.

Language Experts

We recommend Language Experts list their contributions on their CV under a heading for *Web publications*. Language Experts can download their data, and transform them into a PDF form to put on their personal webpage. Download is currently available from the Language page in JSON format. A script is in development to convert this into a Text file.

[34] See Champieux & Coates, this volume and Alperin et al., this volume.

Example:

Web publications

Paul Roger Bassong (2014 -), Language expert on TerraLing, Contributions to the following TerraLing datasets: SSWL (http://terraling.com/groups/7) and Conjunction and Coordination (http://terraling.com/groups/8).

Paul Roger Bassong (2014 -) Basaá - dataset, examples, and comments for SSWL (*Property Values*: 150, examples:151).

Paul Roger Bassong (2017 -), Basaá - dataset, examples, and comments for Conjunction and Disjunction (*Property Values*: 40, examples 42).

Administrators

List your administrative functions under Service to the Field or Reviewing/Editing

Name, [DATE], Administrator of [DATASET].

5. Summary

In this chapter, we provided a general description of the goals, design, structure and potential of TerraLing (http://terraling.com/), as well as a snapshot of its current state in terms of contents. TerraLing is a collection of databases, which is, virtually by definition, constantly in progress, and constantly capable of being enriched and

developed according to the most updated advances in theoretical and comparative linguistics, as well as in digital technologies.

The main purpose of TerraLing is to build a linguistic database of crosslinguistic properties that can support theoretical research. Its basic setup allows working with linguists who are native speakers or signers as language-experts providing the data. This provides researchers with the opportunity to use the tools of theoretical linguistics to access the implicit knowledge of native speakers/signers, in order to probe the cross-linguistic situation. The basic database schema is flexible, which means it can be adapted to the research needs of individual researchers. Since the database codes observable fine-grained variation, the database can in fact support a broad community of scientists. The long-term goal is to turn TerraLing into a ready-made community tool that linguistic projects can use to gather and store their data for comparative research purposes.

References

Bosque, Ignacio, and Violeta Demonte. 1999. *Gramática descriptiva de la lengua española*. Madrid: Colección Nebrija y Bello, Espasa.

Broekhuis, Hans. 2013. *Syntax of Dutch: Adjectives and Adjective phrases*. Amsterdam: Amsterdam University Press. doi:10.26530/OAPEN_431435.

Broekhuis, Hans, and Norbert Corver. 2016. *Syntax of Dutch: Verbs and Verb Phrases – Volume 3*. Amsterdam: Amsterdam University Press.

Broekhuis, Hans, and Marcel den Dikken. 2012. *Syntax of Dutch: Nouns and Noun Phrases – Volume 2.* Amsterdam: Amsterdam University Press. doi:10.26530/OAPEN_431435.

Cinque, Guglielmo. 1994. On the evidence for partial N-movement in the romance DP. In *Paths towards Universal Grammar*, ed. Luigi Rizzi, Raffaella Zanuttini, Jan Koster, and Jean-Yves Pollock, 85–110. Washington, DC: Georgetown University Press.

Cinque, Guglielmo. 2005. Deriving Greenberg's Universal 20 and its exceptions. *Linguistic Inquiry* 36 (3): 315–332.

Cinque, Guglielmo. 2010. *The Syntax of Adjectives: A Comparative Study*. Cambridge: MIT Press. https://doi.org/10.7551/mitpress/9780262014168.001.0001.

Collins, Chris, and Richard Kayne. 2007. A proposal for a database of the syntactic structures of the world's languages. http://ling.auf.net/lingbuzz/003404.

Dixon, Robert M. W. 1982. *Where Have All the Adjectives Gone?* Berlin: Walter de Gruyter.

Dixon, Robert M. W., and Alexandra Y. Aikhenvald, eds. 2004. *Adjective Classes: A Cross-linguistic Typology*, Volume 1. Oxford: Oxford University Press.

Dryer, Matthew S. 2013a. Order of adjective and noun. In *The World Atlas of Language Structures Online*, ed. Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/87. Accessed May 1, 2019.

Dryer, Matthew S. 2013b. Determining dominant word order. In *The World Atlas of Language Structures Online*, ed. Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/s6. Accessed May 19, 2019.

Dryer, Matthew S., and Martin Haspelmath, eds. 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/. Accessed Sept. 28, 2015.

Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful element. In *Universals of Language*, ed. Joseph Greenberg, 73–113. Cambridge, MA: MIT Press.

Holmberg, Anders. 2015. *The Syntax of Yes and No*. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780198701859.001.0001

Huddleston, Rodney, and Geoffrey Pullum. 2005. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Kayne, Richard S. 2013. Comparative syntax. *Lingua* 130: 132–151. https://doi.org/10.1016/j.lingua.2012.10.008.

Schachter, Paul, and Fe T Otanes. 1983. *Tagalog Reference Grammar*. Berkeley: University of California Press.

Schütze, Carson T, and Jon Sprouse. 2014. Judgment data. In *Research Methods in Linguistics*, ed. Robert J. Podesva and Devyani Sharma, 27–50. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139013734.004.

Sportiche, Dominique, Hilda Koopman, and Edward Stabler. 2013. *An Introduction to Syntactic Analysis and Theory*. Hoboken, NJ: John Wiley & Sons.

Sproat, Richard, and Chilin Shih. 1991. The cross-linguistic distribution of adjectival ordering restrictions. In *Interdisciplinary Approaches to Language: Essays in Honor of S.-Y. Kuroda*, ed. C. Georgopoulos and R. Ishihara, 565–593. Dordrecht: Kluwer. doi:10.1007/978-94-011-3818-5_30.

Sprouse, Jon, and Diogo Almeida. 2012. Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics* 48 (3): 609–652. https://doi.org/10.1017/S0022226712000011.

Sprouse, Jon, Carson T Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua* 134: 219–248. https://doi.org/10.1016/j.lingua.2013.07.002.