

TerraLing: the Now and the Future

Hilda Koopman
koopman@ucla.edu
Department of Linguistics, UCLA

Leiden/Bielefeld Workshop on Comparative Syntax,
University of Leiden,
May 23 2024
Panel on Methodology and Resources.

Roadmap

- General Goal of the TerraLing project
- Intro to the TerraLing platform: a (relational) database-backed webapplication
Open-access, open-ended, community-based Koopman et al (2017)
TerraLing, <http://terraling.com/> ¹
- a platform where (individual) researchers can set-up their own research project.
 - The now: Current state
 - How it's set up–The structure of TerraLing datasets and the flexible data-schema
 - *Example*: Towards the future. Brief Report on a pilot on how to construct a syntactic database for Passive-like constructions
(code up what is known, fill in the gaps in the paradigms; generate new datasets for (under)described languages.

¹See Collins & Kayne 2007, Koopman & Guardiano 2022; Dennis Shasha (NYU, Computer Science) as the system architect, Hannan Butt (principal engineer and developer, and Hannan Butt, and Shailesh Vasandani (engineer)

Roadmap

- General Goal of the TerraLing project
- Intro to the TerraLing platform: a (relational) database-backed webapplication
- Open-access, open-ended, community-based Koopman et al (2017)
 TerraLing, <http://terra ling.com/> ¹
- a platform where (individual) researchers can set-up their own research project.
 - The now: Current state
 - How it's set up–The structure of TerraLing datasets and the flexible data-schema
 - *Example:* Towards the future. Brief Report on a pilot on how to construct a syntactic database for Passive-like constructions
 (code up what is known, fill in the gaps in the paradigms; generate new datasets for (under)described languages.
- Differences with existing databases (TerraLing vs. Grambank)
- Challenges!

¹See Collins & Kayne 2007, Koopman & Guardiano 2022; Dennis Shasha (NYU, Computer Science) as the system architect, Hannan Butt (principal engineer and developer, and Hannan Butt, and Shailesh Vasandani (engineer)

TerraLing: general goal

- Theoretical progress demands hypotheses/ specific research questions be tested on the basis of comparative-linguistic data.

urgent and necessary for "survival" of the field

- **Goal of TerraLing**

Construct a tool that can support fundamental theoretical-driven comparative research in (morpho-)syntax and semantics, now and in the future.

*...a long term project, open-access, open-ended,
requires community involvement*

TerraLinging

- Is set up so as to be able to directly involve native speaker linguists
as the theory requires: Native judgments by trained speakers are a crucial tool
- Theories change, and so do technical implementations, terminology

Don't code up the theory. Find a way to generate and code up crosslinguistically comparable empirical data that theories need to account for. Set it up so that known-crosslinguistic variation can be captured. avoid (opaque) linguistic terminology, needs to be strongly decompositional, no binning!

The Now: Current state

- *"en attendant"²... waiting for the implementation of the new user-interface and face-lift (in advanced state of development).*
- <http://terraling.com/>: current datasets and groups.

²"des en attendant", (lit. the while waiting), refers to flipflops in Ivorien French, waiting to afford getting real shoes

Public groups

- SSWL

<https://www.terraling.com/groups/7> -originally started by Chris Collins (with Richie Kayne)); involving various 'teams': *word order* Chris Collins, Andrea Cattaneo, Jim Wood, Hilda Koopman, Cristina Guardiano *negation* Vesilanova, Koopman *polar questions and answers* (Anders Holmberg, Craig Sailor), *Bare nouns and (in)definite articles* (Cristina Guardiano and Hilda Koopman)

Public groups

- SSWL

<https://www.terraling.com/groups/7> -originally started by Chris Collins (with Richie Kayne)); involving various 'teams': *word order* Chris Collins, Andrea Cattaneo, Jim Wood, Hilda Koopman, Cristina Guardiano *negation* Vesilanova, *Koopman polar questions and answers* (Anders Holmberg, Craig Sailor), *Bare nouns and (in)definite articles* (Cristina Guardiano and Hilda Koopman)

- Cinque's Universal 20 database

<https://www.terraling.com/groups/15>

- Passive-like constructions (pilot)

<https://www.terraling.com/groups/13> Hilda Koopman.

- Coordination and Conjunction.

<https://www.terraling.com/groups/8>

<https://sites.google.com/view/viola-schmitt/current-projects>

- Quantification and Plurality.

<https://www.terraling.com/groups/20>

<https://sites.google.com/view/viola-schmitt/current-projects>

Private groups

- Subject properties ((in)definite ...)
Cristina Guardiano and Hilda Koopman
- Anaphora Isabelle Charnavel and Dominique Sportiche.
- The syntax and semantics of (in)definites and bare nouns
(based on lessons learned from the pilot in SSWL- Cristina Guardiano and Hilda Koopman)
- Definiteness and Genericity. (Aviv Schonfield and Magdalena Roszkowski)

The structure of TerraLing datasets

- Each TerraLing dataset is constructed around an open-ended set of **linguistic entities**.

The structure of TerraLing datasets

- Each TerraLing dataset is constructed around an open-ended set of **linguistic entities**.
- **Linguistic entities** can be anything, languages, constructions, sentences or even individual lexical items.

SSWL these are "languages", whether spoken, signed, extinct or emerging, dialects, or grammars of an individual speaker.

The structure of TerraLing datasets

- Each TerraLing dataset is constructed around an open-ended set of **linguistic entities**.
- **Linguistic entities** can be anything, languages, constructions, sentences or even individual lexical items.

SSWL these are "languages", whether spoken, signed, extinct or emerging, dialects, or grammars of an individual speaker.

- **Properties, property values and examples.** Each TerraLing dataset contains a set of so-called properties that serve to classify and compare the basic linguistic entities.

The structure of TerraLing datasets

- Each TerraLing dataset is constructed around an open-ended set of **linguistic entities**.
- **Linguistic entities** can be anything, languages, constructions, sentences or even individual lexical items.

SSWL these are "languages", whether spoken, signed, extinct or emerging, dialects, or grammars of an individual speaker.

- **Properties, property values and examples.** Each TerraLing dataset contains a set of so-called properties that serve to classify and compare the basic linguistic entities.
- A **property** is essentially a yes-no question that can be answered for each linguistic entity in the dataset. Each property: detailed description; detailed instructions to construct the relevant examples; and how its value is to be determined for a given linguistic entity (**values: yes/no/NA/U(nknown)**).






Word order example from SSWL

- word order– no notion of dominant order. WYSIWYG
- 15_Num N
- 16_N Num

Search results

SSWL Search Results

Repeat search with same keywords

Views    Actions  

| Property Name | Property Value | Property Name | Property Value | Count |
|---------------|----------------|---------------|----------------|-------|
| 16_N Num | No | 15_Num N | Yes | 108 |
| 16_N Num | Yes | 15_Num N | No | 87 |
| 16_N Num | Yes | 15_Num N | Yes | 52 |
| 16_N Num | No | 15_Num N | No | 0 |

- 52 languages/256: both orders NumN and NNum.
 - Variables of variation → Definiteness, type of numeral,.....
 - Definiteness. Theory?
 - N (or rather remnant NP) moves leftwards into D region above Num. $D > \text{Num}$)
- Theoretical expectation: Gaps! Expected: No language with *Num
 N = definite and N Num = indefinite.

One-level datasets, and two level datasets

one level dataset

Languages– Prop of Language

ex: SSWL

two level dataset:

Languages– Prop of Language

Forms – Properties of forms

ex.: Universals and Plurality-
Passive-like constructions (pilot)

A two-level dataset: Languages and Forms

- How to: – Passive-like constructions (pilot)

<https://www.terraling.com/groups/13>.

- ...heavily studied, huge body of knowledge, prime example for modular syntactic accounts.

Types "e.g. constructions" and Forms..

- E: Canonical Aux-Part Passives (Agent Theme Vs), *but also get passives, adjectival passives, Middles, anticausatives, pseudo-passives, easy-to-please constructions, -able(potentials).*
- Other languages: also impersonal passives, long passives, malefactive, indirect passives, ...
- **Forms** vary -
 - language internally: more than one type with different forms: Italian -si passives, and Aux-participle passives *each with further subsets*
 - *from* "same" forms: (English: adjectival passive and verbal passives).
 - *all the way to* "same" form for many different constructions (Japanese -(r)are).
 - or a subset of some of the constructions. (Bantu: different forms for stative (=E. adjectival passives) and passive).
 - Not one form = a single syntactic environment.

but one (invariant) -(r)are with distributional differences following from height of Merge? (Cinque (2022), Ishizuka and Koopman, in prep)

A toy example from English: Table of variation

| "Types" → | AdjPass | VerbalPass | Middle | EasytoPlease | -able |
|--------------------------|-------------|------------|--------|--------------|-------|
| Forms → | (be) + Part | (be)+Part | V | to V | -able |
| Prop. Forms ↓ | | | | | |
| 1.Act/Pass Dist on V(s)? | Y | Y | N | N | N |

- Hypothesis for English: Passive Voice is zero? (Koopman, 2021)

A toy example from English: Table of variation

| "Types" → | AdjPass | VerbalPass | Middle | EasytoPlease | -able | |
|----------------------------------|-------------|------------|--------|--------------|-------|--|
| Forms → | (be) + Part | (be)+Part | V | to V | -able | |
| Prop Forms ↓ | | | | | | |
| 1.Act/Pass Dist on V(s)? | Y | Y | N | N | N | |
| Type of pred | | | | | | |
| (simple) transitive | | | | | | |
| 4.agent>theme | Y | Y | Y | Y | Y | |
| 6.cause _{inanim} >theme | Y | Y | Y | N | ?Y | |
| 5.experiencer>theme | Y | Y | N | ?N | ?Y | |
| 5a. theme-experiencer | ? | Y | N | N | N | |
| intransitive | | | | | | |
| 7.(simple) unacc | N | N | N | N | N | |
| 8.(simple) unerg | N | N | N | ?N | ?N | |
| double object | | | | | | |
| etc.. | | | | | | |

A (partial) Table of Cross-linguistic variation

| "Languages"→ | English | E.Middle | Jamaican | Mandarin | Samoan | Ewe |
|----------------------------|-----------|----------|----------|----------|--------|-----|
| Forms → | (be) Part | "wash" | V | V.le? | V | V |
| Properties Forms↓ | | | | | | |
| 1.Act/Pass Dist on V(s)? | Y | N | N | N | N | N |
| 25. Implicit ext.argument? | Y | Y | Y | Y? | Y? | N |
| 13. by-phrase? | Y | N | N | N | Y? | N |
| 20. promotion? | Y | Y | Y | Y | ? | N |

- Comment: orange: distinguishing between pro drop, and other types of available interpretations needs to be further investigated. (?, ?)
- Many languages: Passive Voice is zero? (Koopman, 2021)

A more fine-grained typology

- G1: A Language with no canonical passive construction (Ewe, Kwa, Niger Congo) is a language that has a NO value for this set of properties.
- On languages that have canonical eventive passives with no change of form between active and passive Vs : (see Keenan & Dryer (2007), Roberts (2019), <https://apics-online.info/parameters/902/30.3/10.0>. Koopman (2012).)
- row 1: bears on the presence of silent elements in the syntax and the "signature" they leave. →(Can) Passive Voice be silent? ³

³Such cases are surely underreported in literature. (See Keenan & Dryer (2007), Roberts (2019) for list of languages. Cf. Ken Hale, as cited in ? on Australian Languages, and how the only correlation between ergative and accusative languages is the absence of a morphological distinction between active and passive verbs in ergative languages.

Form-based properties: Advantages

- Forms and their distributions are ultimately the empirical picture theories must capture.
- Direct connection to the forms, and their distributional properties within a language, and across languages
 - Gets around definitions "what is passive, a middle, an anticausative, a potential, a stative etc; the unreliability of glosses, and the problem of "dialects" of linguistic terminology (tower of Babel)".
 - Can capture finer distinctions (differences between strong and weak forms!)
 - Homophony: Allows refining the forms over time. (from a single *-able* construction to two types of *-able* constructions (able1 and able 2), from Adjectival Passives to two (or more) types of Adjectival passives, etc).
 - allows coding up multiple forms used in passive-like constructions in a language.
- Allows native speaker linguists of un(der)-described languages or fieldworkers to participate, and collaborate.

TerraLing vs. Grambank

| | TerraLing | Grambank |
|--------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|-----------------------------------------------------------------|
| mode of data generation | "fieldmethods" (+ grammar/publ. based) +direct involvement of native speaker linguists | grammar based (no/little) involvement of native speakers |
| property based coding schema: <i>but</i> | yes/no question binary features | yes-no questions went from Wals to binary features |
| support fundamental research in syntactic theory? in semantics typology/"computational" linguistics | yes, in principle yes yes, in principle | no yes, huge! |
| community based | small core! | yes, very large! |

- Grambank. great user-interface, and advice:
<https://github.com/grambank/grambank/wiki/Advice-for-typological-database-construction>

but...Grambank properties

- **Unsuitable for theoretical investigation...** Check the description of the properties and the coding schema and the sources.
- Grambank: GB020. Are there definite or specific articles?
- An article is a marker that accompanies the noun and expresses notions such as (non-)specificity and (in)definiteness. Sometimes these notions of specificity and definiteness are summed up in the term 'identifiability'.
 - The formal expression is irrelevant; articles can be free, bound, or marked by suprasegmental markers such as tone. *Agreed!*
- *Articles are different from demonstratives in that demonstratives occur in a paradigm of markers that have a clear spatial deictic function. As demonstratives can grammaticalise into definite or specific articles, they form a natural continuum, making it hard to define discrete categories, but to qualify as an article a marker should be used in some cases to express definiteness without also expressing a spatial deictic meaning.*
- Procedure
 - Code 1 if there is a morpheme that can mark definiteness or specificity without also conveying a spatial deictic meaning.
 - Code 0 if the source does not mention a definite article and you cannot find one in examples or texts in an otherwise comprehensive grammar.
 - Code ? if the grammar does not contain enough analysis to determine whether there is a definite article or not.

Challenges: from "en attendant" to the future

- Money! Money! Money! User interface needs to be implemented to bring the project to the next level.
- How can you get involved?? (Wo)manpower, there is a ton to do! (through workshops, working groups, meetings, using it for your own comparative research project, etc?).
- How to get the (syntactic) community more actively involved! how to achieve this? All kinds of ideas, but...

- Interest of theoretical syntacticians (in the US) has been low so far—?

Perhaps, the specific Minimalist Program Path that syntax has taken, where Minimalist=structural economy (i.e. *what can be removed from the syntax, where can we shift it to*), is incompatible with the specific path to reach the goal of this project, (which is more like a "linguistics genomics" program), requiring comparative work at a massive and finely grained scale (strongly decompositional, and Merge based with very simple general constraints, looking for gaps, predictions etc).

- Get Involved!! With your friends, students, colleagues, etc..

Thank you!!

- to the TerraLing team: to Dennis Shasha, architect of the TerraLing, and the development team Hannan Butt (principal engineer), Shailesh Vasandani (engineer).
- ...and the members of the TerraLing community, in particular to Nina Haslinger and Viola Schmitt, to Nikos Angelopoulos, András Bárany, Paul Roger Bassong, Guglielmo Cinque, Cristina Guardiano, Jutta Hartmann, JosuÓ Henoc, Vincent Homer, Travis Major, Victoria Matthieu, Pamela Munro, Harold Torrence, Augustina Owuso, Ethan Poole, Magdalena Roszkowski, Valerie Wurm, Gert Jan Postma.
- ... And a special thanks to the participants of different fieldmethods classes at UCLA over the years, to the SSWL workshop participants at ALS, Cote d Ivoire, Université de Youndé 2 (Cameroon), and ILA (Institut de Linguistique Appliqué (Abidjan)).
- Get involved!! Talk to your colleagues, friends, sign up for our zoom meetings. Plenty of ways to get involved. If you have datasets, think of putting them up on TerraLing, If you or your students want to set up a dataset, or help develop one, collaborate! If you want to participate with a group, do so: come to our meetings! (Starting up again in the Fall) . Help develop questionnaires. Develop table of variations for subsections etc.... Ask for grant money!
- to Sjef, András and Jutta for organising this wonderful workshop!

Some selected References

- Cinque, Giglielmo. 2022. Different Voice heads. A view from long-passivization. Presented at the second Terraling workshop.
- Collins, Chris & Richard Kayne. 2007. A Proposal for a Database of the Syntactic Structures of the World's Languages. <http://ling.auf.net/lingbuzz/003404>.
- Keenan, Edward L & Matthew S Dryer. 2007. Passive in the world's languages. In Timothy Shopen (ed.) (ed.), Language typology and syntactic description, vol. I. Clause structure., 325–361. Cambridge: Cambridge University Press.
- Koopman, Hilda. 2012. Samoan ergativity as double passivization. In Brugé et al (ed.), Functional Heads: The Cartography of Syntactic Structures, vol. 7, 168–180. Oxford University Press.
- Koopman, Hilda & Cristina Guardiano. 2022. Managing Data in TerraLing, a Large-Scale Cross-Linguistic Database of Morphological, Syntactic, and Semantic Patterns. In The Open Handbook of Linguistic Data Management, The MIT Press. doi:10.7551/mitpress/12200.003.0060.
- Roberts, Ian. 2019. Parameter Hierarchies and Universal grammar. Oxford University Press.