CLLD apps as a tool for the construction of datasets

Sascha Alexeyenko

University of Göttingen

Workshop on Cross-Linguistic Data Formats

Leipzig — 14 Dec 2023

A common use case

- There is an **existing** dataset
- Its creator wants to make a CLDF-conform database out of it
- Some magic happens, a.k.a. @xrotwang
- The database gets created (+served by a CLLD app, optionally)

Today, I will discuss a different scenario in which the CLDF-CLLD architecture is used as a tool **during** the creation of a dataset.

Plan for today

- I will first introduce the Nominal Modification Database (NMDB), which I am constructing since several years
- In parallel, I will discuss my experiences with using CLDF-CLLD as a organization + visualization tool for the data
- And I will share some ideas about what could potentially be modified in the default CLLD app in order to make it more friendly for the researchers with a similar use case to mine

Origins of the idea

- Joint work with Hedde Zeijlstra on the **Head-Final Filter** (i.e. the ban on intervening dependents in prenominal modifiers in languages like English and German) made us check how languages beyond SAE deal with this constraint.

*a proud of his son father	*ein stolzer auf seinen Sohn Vater
a father proud of his son	ein auf seinen Sohn stolzer Vater

- In general, the nominal domain tends to be a poor sister of the verbal/clausal domain as far as the availability of broad typological information is concerned (with the exception of things like Dem-Num-Adj-N orders studied in connection with Greenberg’s Universal 20).

The goal of the database

Systematic information about:

- the availability of various kinds of nominal modifiers in the languages of the world
- the possibility for them to be complex
- the word orders available to them
- their marking (at the moment: info about “linkers/attributivizers”; in the prospect: details about their φ-agreement)

Data collection

- In 99% of the cases, the relevant information is not available in the descriptive grammars
- For this reason, we are forced to look for linguists specialized in the language in question, who may have contacts with native speakers or have the relevant information in their fieldwork notes, corpora, or other resources
- It is thus a rather slow and painful data collection process, most of which happens in the form of an email exchange

Data collection

- the long-term nature of the data collection process
- the scatteredness of the obtained data across places

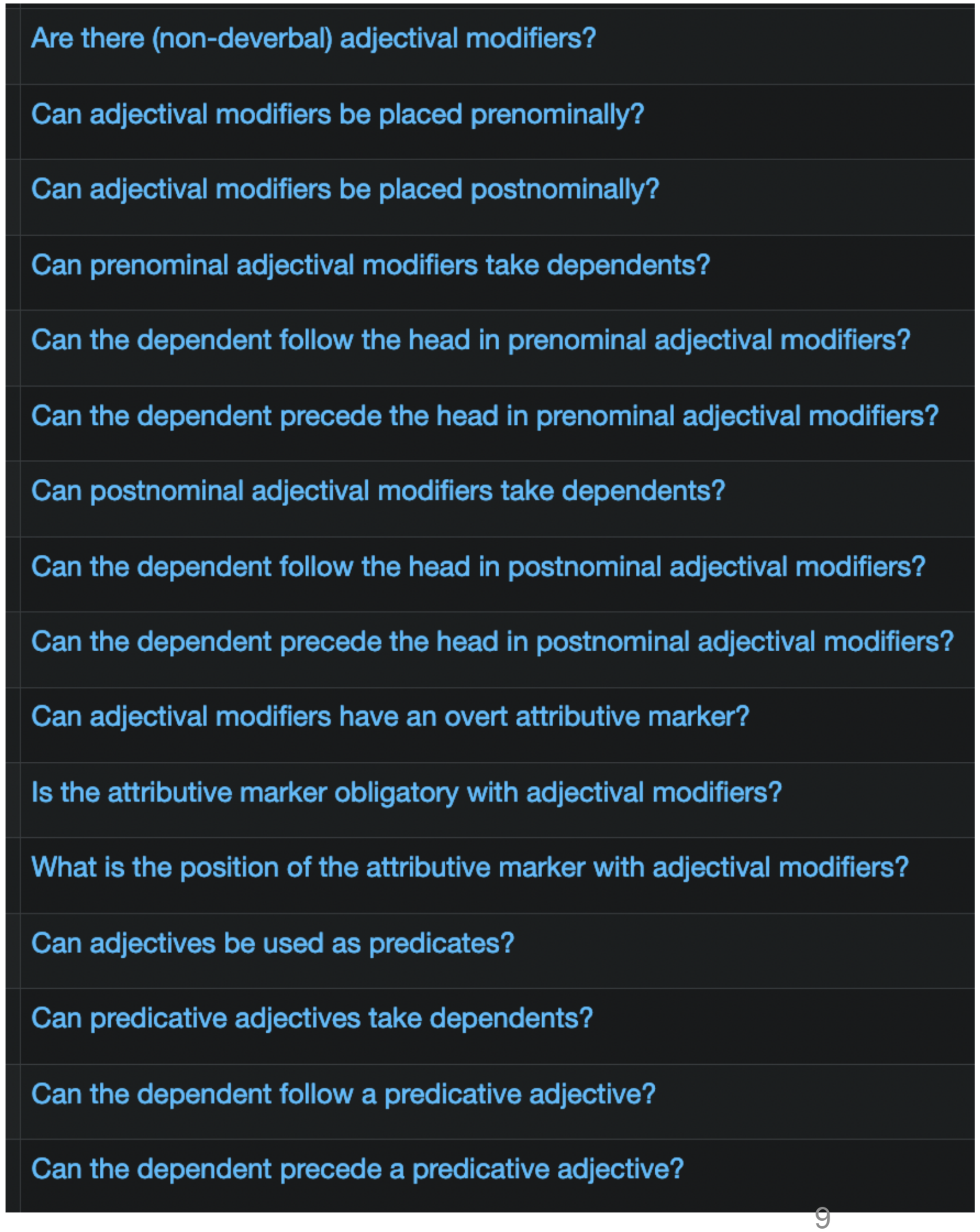
made it necessary to use some structured database platform as an organization and visualization tool

The structure of NMDB

1. parameters
2. languages
3. values
4. examples

The structure of NMDB

1. parameters
   - at the moment 50, subject to change
   - cover common adnominal modifiers
     - adjective-like items
     - deverbal/clausal modifiers
     - adnominal adpositional phrases



Are there (non-deverbal) adjectival modifiers?

Can adjectival modifiers be placed prenominally?

Can adjectival modifiers be placed postnominally?

Can prenominal adjectival modifiers take dependents?

Can the dependent follow the head in prenominal adjectival modifiers?

Can the dependent precede the head in prenominal adjectival modifiers?

Can postnominal adjectival modifiers take dependents?

Can the dependent follow the head in postnominal adjectival modifiers?

Can the dependent precede the head in postnominal adjectival modifiers?

Can adjectival modifiers have an overt attributive marker?

Is the attributive marker obligatory with adjectival modifiers?

What is the position of the attributive marker with adjectival modifiers?

Can adjectives be used as predicates?

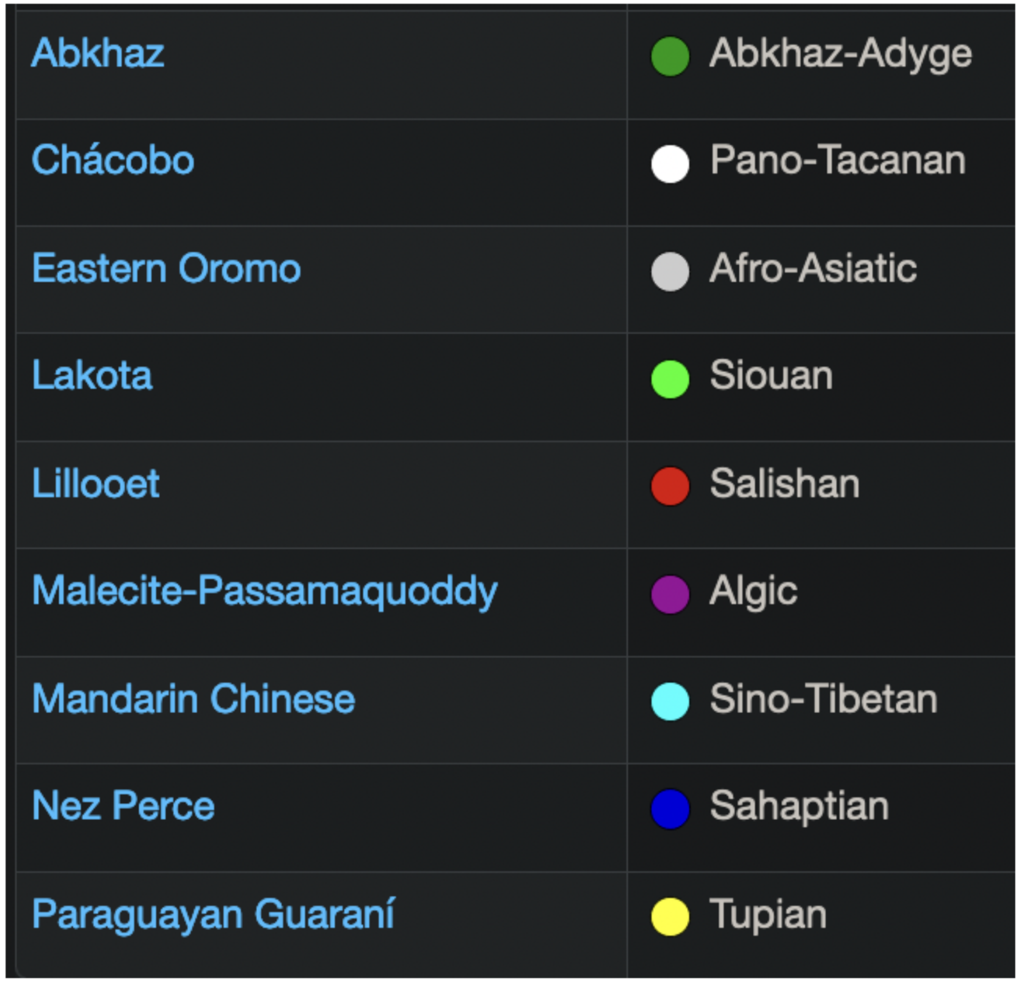
Can predicative adjectives take dependents?

Can the dependent follow a predicative adjective?

Can the dependent precede a predicative adjective?

The structure of NMDB

2. languages
   - goal: (min) one language per family (Glottolog classification)
   - at the moment: 10 languages are in the database, data from ~10 lngs will be added in the next time



Abkhaz	● Abkhaz-Adyge
Chácobo	● Pano-Tacanan
Eastern Oromo	● Afro-Asiatic
Lakota	● Siouan
Lillooet	● Salishan
Malecite-Passamaquoddy	● Algic
Mandarin Chinese	● Sino-Tibetan
Nez Perce	● Sahaptian
Paraguayan Guaraní	● Tupian

The structure of NMDB

3. values
    - quasi-binary (‘yes/no’, ‘yes/no but’, ‘unknown’, ‘undefined’)
4. examples
    - almost for every value
    - grammatical and ungrammatical ones
    - the central point of the database

Discussion

- I presume that the use case I have been describing here is actually a very common one among linguists, and many would be happy to make use of the CLDF-CLLD architecture
- However, the burden of having to do the initial setup of the system may be hard/disencouraging for many linguists
- From my side, I'd be happy to contribute to the documentation based on the notes I have been making
- On the other hand, some features of the default setup may be modified to make the start easier for a linguist
  - Glottolog integration
  - incorporation of glossed examples

Thank you for your attention!

Thanks also go to:

- Natasha Thalluri, Tereza Zikudová, Ya Gao (for help with data collection)
- Robert Forkel (for obvious reasons)
- Hedde Zeijlstra (for general support)