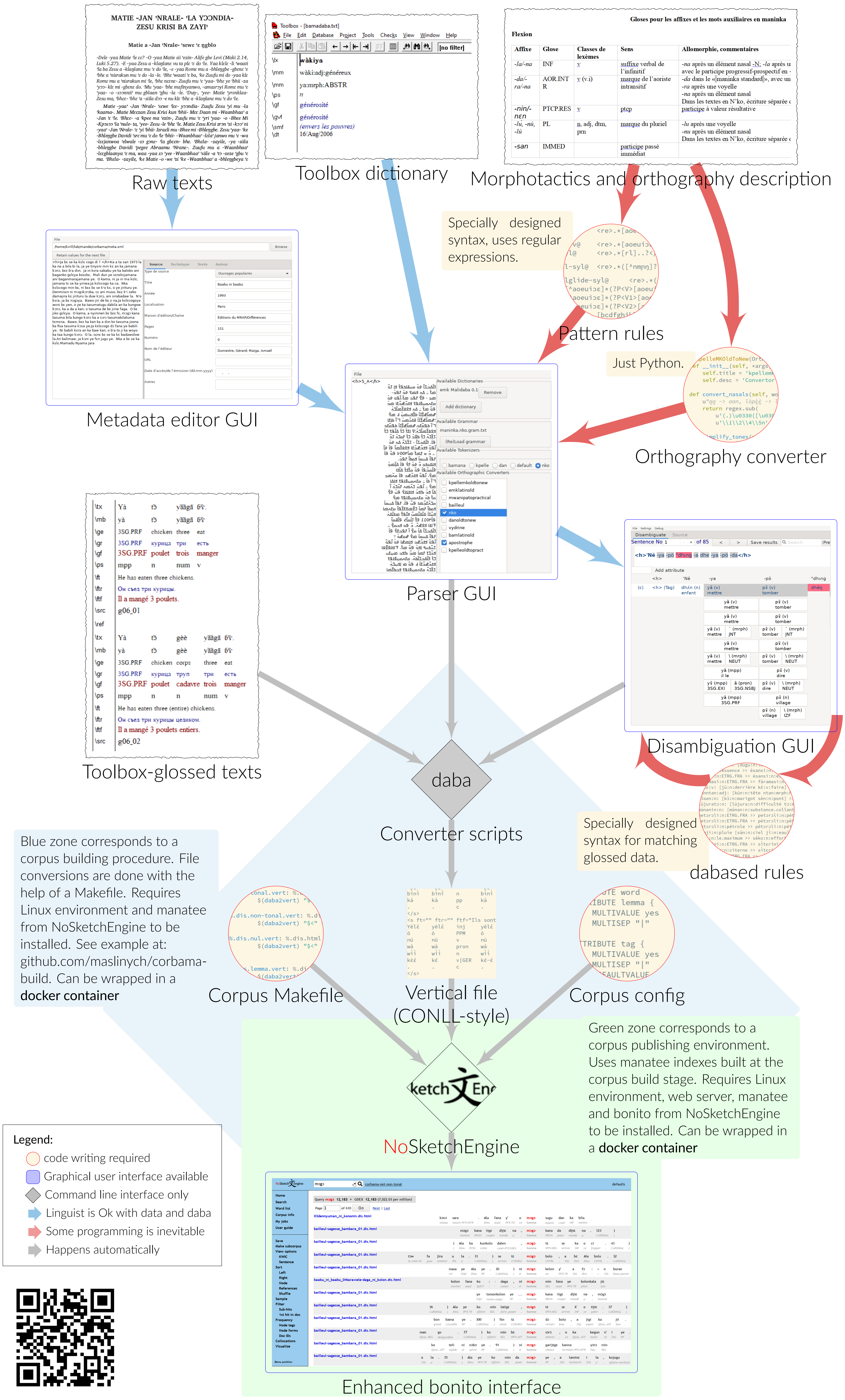
Daba software for written corpora of underresourced languages

Kirill Maslinsky [1,2] Valentin Vydrin [3,4]

[1]Higher School of Economics, St. Petersburg, Russia [2]Institute of Russian Literature of Russsian Academy of Sciences, St. Petersburg, Russia
[3]Inalco, Paris, France [4]LLACAN (CNRS, UMR-8135), Paris, France

Building a corpus with daba: A typical workflow



Blue zone corresponds to a corpus building procedure. File conversions are done with the help of a Makefile. Requires Linux environment and manatee from NoSketchEngine to be installed. See example at: github.com/maslinych/corbama-build. Can be wrapped in a **docker container**

Green zone corresponds to a corpus publishing environment. Uses manatee indexes built at the corpus build stage. Requires Linux environment, web server, manatee and bonito from NoSketchEngine to be installed. Can be wrapped in a **docker container**

Legend:
- ○ code writing required
- ▭ Graphical user interface available
- ◆ Command line interface only
- ➡ Linguist is Ok with data and daba
- ➡ Some programming is inevitable
- ➡ Happens automatically

Software implementation

- **Cross-platform**: Linux, Windows, MacOS
- **Python** 2.7, 3; **wxPython** GUI library
- **Open source**: github.com/maslinych/daba

Corpora built with daba

Corpora of Mande languages built as a part of Corpora Mandeica project: http://cormand.huma-num.fr/mandeica/

- Corpus Bambara de Référence (11M)
- Corpus Maninka de Référence (3.5M)
- Corpus du dan de l'Est (0.5M)
- Corpus Mwan (0.05M)
- ...further corpora in progress

Daba's habitat

- limited amount of texts available, especially in electronic formats. As a rule, such languages are underrepresented on the Internet;
- low level of standardization, sometimes competing orthographic systems;
- often insufficient grammatical and lexicographical description;
- corpora built and used mostly by non-native speakers (linguists, language learners etc.) — glosses in a European language required.

Daba features

- A framework to create **rule-based morphological parsers** — allows bootstrapping corpora without annotated training data.
- Support for **hierarchically glossed texts**.
- Graphical user **interface for manual disambiguation**.
- **Integration with legacy formats** ubiquitous in field linguists' data — SIL Toolbox.

Daba caveats

- **A Linguist + Daba ≠ Corpus.** Software developer is still required.
- **Each corpus is different** and requires writing new code: pattern rules, orthographic converters, Makefile for building a corpus etc..
- Work required to publish a smaller corpus is typically **larger** due to even more inconsistency and instability of writing and linguistic description characteristic of underresourced languages.

https://github.com/maslinych/daba LIFT 2019, Orléan kmaslinsky@hse.ru vydrine@gmail.com