Linguistic Type Database Update

Francis **Bond**,
Michael Wayne Goodman and Luís Morgado da Costa

Department of Asian Studies,
Palacký University, Olomouc, Czechia

<bond@ieee.org>

DELPH-IN 2022

Faculty
of Arts

Why the LTDB?

- It is hard to work on a grammar that you did not write (just like any software)
  - Or that you wrote in collaboration
  - Or that you wrote sometime ago
  - Or that is generated by the MATRIX
- It is hard to be consistent within a treebank
  - Especially if it has multiple annotators
  - Or that you treebanked some bits some time ago
  - Or just if it is very big
- LTDB is an attempt to store the information you had in your mind when you wrote the grammar and make it more accessible
  - inspired by literate programming (Knuth, 1992)
  - store documentation **about the grammar** — **in the grammar files**

Faculty
of Arts

LTDB

- Completely rewritten Lexical Type Database (Hashimoto et al., 2007a,b)
- Generalized in 2014 to handle all types (and some instances)

The Linguistic Type Database

status	thing	source	endi
type	normal type		
ltype	lexical type	(type and in lexicon)	lt
rule	grammar rule	(LKB::*RULES)	c
lrule	lexical rule	(LKB::*LRULES)	
irule	inflectional rule	(LKB::*LRULES and (inflectional-rule-p id))	
root	start symbol	(LKB::*root-entries*)	

Rules also list number of daughters and head daughter.

Faculty of Arts

Headedness

We are **Head-driven** Phrase Structure Grammar, so it is nice to know the headedness of rules. We record 5 different possibilities:

- ▲ unary: headed
- △ unary: non-headed
- ◭ binary: left-headed
- ◮ binary: right-headed
- ◬ binary: non-headed

For each rule, in look for the daughters of the rule, see if `*head-daughter-path*` exists (only implemented for LKB at the moment).

Use the new-ish comment field

Originally:

```
; <type val="n_-_c_le">
; <description>Intransitive count noun (icn)
; <ex>The dog barked.
; <nex>
; <todo>
; </type>
n_-_c_le := n_intr_lex_entry.
```

This becomes (`ltype-comment`):

```
n_-_c_le := n_intr_lex_entry
"""Intransitive count noun (icn)
<ex>The dog barked.
<nex>Much dog bark.""".
```

Faculty
of Arts

Other Changes

- Integration with grammar catalogue
- Description written in Restructured Text
  - Allows more flexible formatting
  - Special macros for positive and negative examples
- Scripts written in python3
- Source available in github:
  `https://github.com/fcbond/ltdb`

2020 enhancements

- ACE, LKB and PET now allow docstrings with `"""  """` on all types and instances, to read them all
    - Thanks everyone for their support.
- The fftb can link to this for rules and lexical types
    - Maybe we should include an LTDB url in the metadata
- Moved to python3
- Now read tdl with PyDelphin
- You can specify a particular grammar (script file or ace config)
  latest version a branch on github, will move next week

Faculty
of Arts

2022 Enhancements

- Trees and MRS displayed using javascript (like viz-demo)
- Search for MRS predicates in the corpus, as well as types and words
- Slightly more robust
- Can read grammars with LKB (using `lkb/script`) or PyDelphin (using `ace/config.tdl`) or both
- Can pre-load some lisp before reading the config file e.g. to load the `mal` grammar:

```
./make-ltdb.bash --lisp '(push :mal *features*)'
--script /path/to/grammar/lkb/script
--acecfg /path/to/grammar/ace/config-mal.tdl
```

Faculty
of Arts

Other useful information

- Make the conversion logs available (so the grammar developer or user can see if there are any known issues) — typically not all MRS’s can be converted to DRMS or JSON
- Give a link to a compressed version of the database, so people can download it — may be easier to access the trees and MRSs for non–delph-in users
  there have been issues with people failing to get MRSs int he past, …

Faculty of Arts

Discussion I

- Who is using this?
- Any requests?
- We will try to host ltdb
- It could interface with fftb better
- Can the matrix add doc-strings?
  - 232 types, some more features
  - 10 grammarians could probably do it in an hour, …
  - Can we do it now, while we have so many knowledgeable matrix people, …?
  - Can matrix libraries add doc-strings for new types?
- Should we attempt to add links to other ontologies such as GOLD or the Norwegian subcat lexicon

Faculty of Arts

Discussion II

- If we want to annotated features (like INFLECTED of MC), where should this go? In the type that first introduces them? Is there a way to index this (i.e. can we output it automatically from the lkb or pydelph<in)?
- Still need help from John to get doc-strings for lexicons, ...
- Should link the examples linked to test framework

Faculty of Arts

References I

Chikara Hashimoto, Francis Bond, and Dan Flickinger. 2007a. The lextype DB: A web-based framework for collaborative multilingual grammar and treebank development. In *First International Workshop on Intercultural Collaboration (IWIC-2007)*, pages 44–58.

Chikara Hashimoto, Francis Bond, Takaaki Tanaka, and Melanie Siegel. 2007b. Semi-automatic documentation of an implemented linguistic grammar augmented with a treebank. *Language Resources and Evaluation*, 42(2):117–126. URL http://dx.doi.org/10.1007/s10579-008-9065-9, (Special issue on Asian language technology).

Donald E. Knuth. 1992. *Literate Programming*. CSLI Publications.

Faculty of Arts