LKB-FOS Update

John Carroll
University of Sussex, UK

DELPH-IN Summit, June 2023

Outline

New features: YY input mode, post-generation chart mapping, parsing efficiency tools, type docstring tooltips

Other improvements: chart dependencies reimplementation, refactoring, new parameters, bug fixes

Some stuff I don't understand, 'todo' list

New Features

YY input mode

A way of integrating external preprocessing modules (tokenisation, tagging, morphological analysis, etc.) into the analysis pipeline

Based on description at <https://github.com/delph-in/docs/wiki/PetInput\#yy-input-mode>

To use it:

1. set parameter `*yy-application*` to name of unix binary / shell script that reads a sentence (on a single line) and outputs a (one-line) YY representation
2. call `initialize-yy`, which starts up this preprocessing application

Application must stream data without buffering (i.e. Python `-u`, FreeLing `--flush`)

Used in latest version of the Spanish Resource Grammar to invoke FreeLing

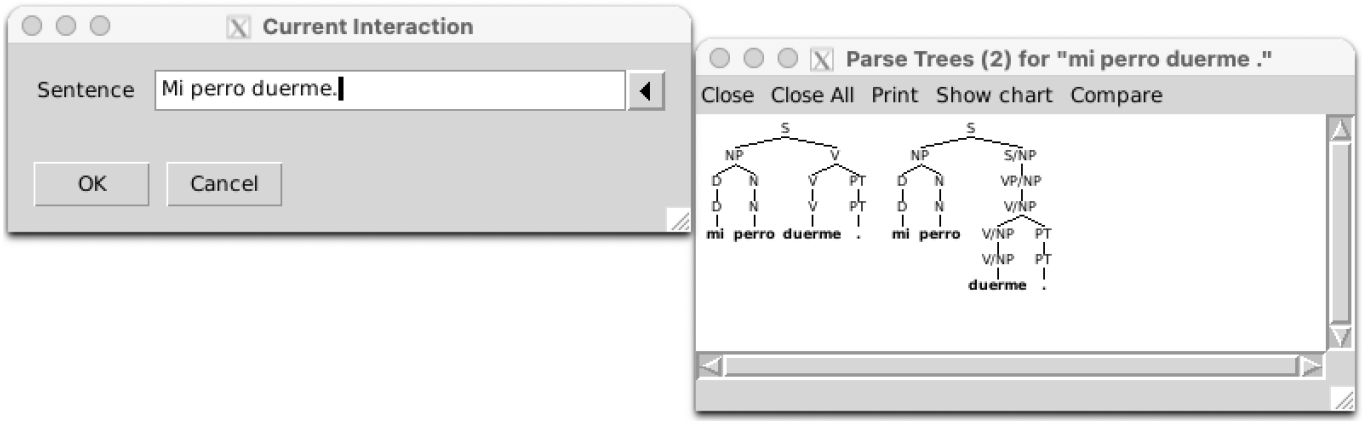
lkb/globals.lsp

```
(defparameter *yy-application*
  #+:linux "bash ~/srg/util/srg-yy-python-api.sh"
  #-:linux "bash ~/srg/util/srg-yy.sh")
```

lkb/script

```
(initialize-yy)
```

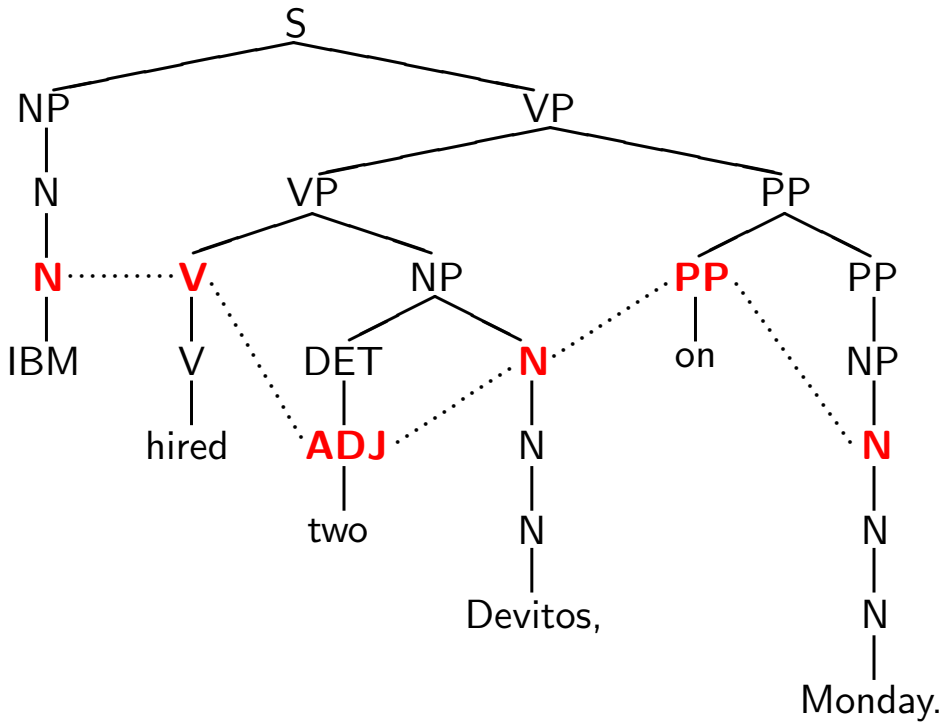
Completely transparent to the LKB user



Post-generation chart mapping

PG mapping rules act on lattices extracted from generation results. Rules loaded with function `read-post-generation-mapping-file-aux`

NB to provide correct `ORTH` feature values, edges in lattice are not necessarily leaves of the generator result tree



Parsing efficiency tools

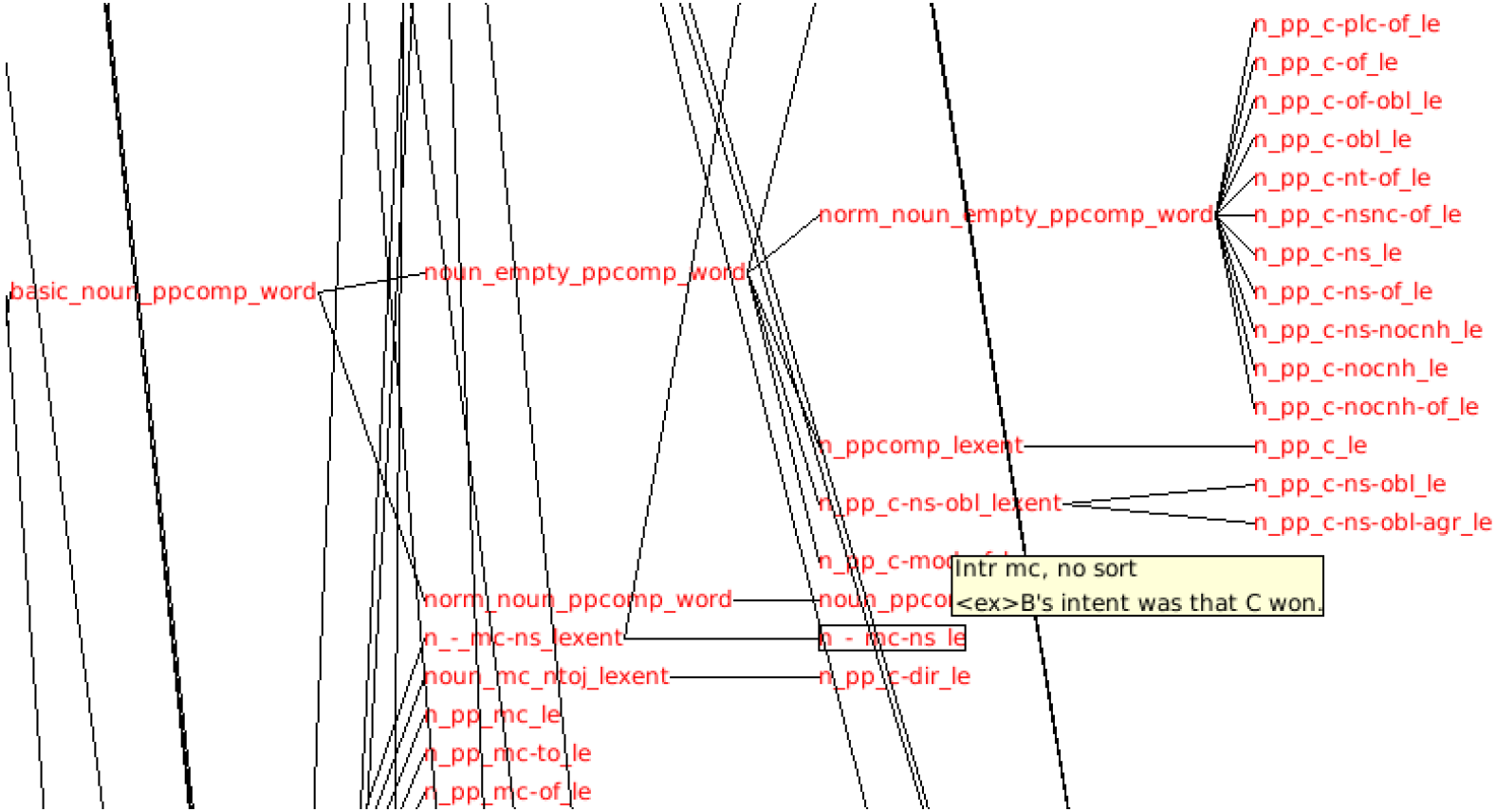
```
(print-chart-counts)
Total cell counts: 336
     end  1    2    3    4    5    6
start
    0     6   19   24   10    8   13
    1         64   27   18   12   18
    2              31    9    0    6
    3                   31    1    4
    4                        11   10
    5                             14
```

IBM hired two Devitos, on Monday.
0 1 2 3 4 5 6

```
(check-subsume-edges 362 371)
Subsumption not forward between PAST_OR_SUBJ_TAM and TAM
  at < SYNSEM : LOCAL : CAT : HEAD : TAM >
Subsumption not forward between NONPRESENT and BASIC_TENSE
  at < SYNSEM : LOCAL : CAT : HEAD : TAM : TENSE >
Subsumption not backward between BASIC_ASPECT and ASPECT
  at < SYNSEM : LOCAL : CAT : HEAD : TAM : ASPECT >
Subsumption check failed in both directions
```

Type docstring tooltips

Hover mouse over a type name – in a type hierarchy or (LKB-native) FS window



Other Improvements

Chart dependencies

E.g.

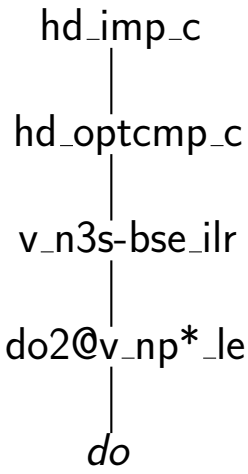
```
... (SYNSEM LKEYS --+COMPKEY) (SYNSEM LOCAL CAT HEAD MINORS MIN)
```

Re-implemented chart dependencies ('chart reduction') code, having noticed two things seemingly wrong with previous implementation:

1. when checking for edges that fulfil chart dependency requirements, we should look only in *other* chart cells and *not* in the current cell (i.e. to filter lexical edge x based on the presence/absence of a selected-for argument y, we should not look for y in the same chart cell as x)
2. if we find that edge x is the only edge fulfilling the requirements of edge y but later we filter out x, then y should also be filtered out; this implies we must iterate until fixpoint is reached

Previous situation with ERG in LKB (and ACE?)

```
0-1 [18] DO2 => (do) <15>
0-1 [53] V_AUX-TAG_DLR => (do) <52>
0-1 [54] V_AUX-NEG-ELL_DLR => (do) <51>
```



Edge 54 dependencies satisfied by edge 53
Edge 53 dependencies satisfied by edge 54
Edge 18 dependencies satisfied by edge 54

But these edges are all in the same chart cell! If condition 1 is applied, edges 18 and 54 are both filtered out. Edge 18 should not be filtered, so perhaps the dependencies are wrong?

As a temporary expedient, condition 1 is currently disabled in LKB-FOS

Internal improvements, comparisons

Refactored code and refined algorithms in core unification functions, the parser and quickcheck

→ up to 25% improvement in parse time

Parsing Rondane with ERG 2018, chart mapping but no PoS tagging, computing top-ranked parse, resource limits giving ~25 timeouts

Mac Intel	*mm:ss*	*Mac M1*	*mm:ss*
LKB-FOS	20:50	LKB-FOS native	10:56
		LKB-FOS emulated	11:45
ACE	14:35	ACE emulated	10:21
+ --disable-generalization	26.51	+ --disable-generalization	17:56

New and changed parameters, bugs and fixes

- `*non-empty-list-type*` no longer used; instead inferred via list head and tail features
- added variable `*cd-debug*` to help debug chart dependencies
- morphological generation smarter about choice of upper-/lowercase in the output
- moved application of chart-dependencies from before lexical parsing to after (in line with other DELPH-IN processors)
- fixed bit rot to get the SPPP input format working again
- fixed a bug where the last part of a sentence was ignored if none of the tokens in it had any morphological analyses

There's some stuff in generation I don't fully understand – especially MRS subsumption and MRS variable property mapping

Summary

Development has continued over the past year

- new features (YY mode, post-generation mapping, efficiency tools, tooltips)
- reimplementation and refactoring
- bug fixes

Still a long 'todo' list, including

- remove default unification and passive chart parser
- add 'grandparent' features in selective unpacking
- unified grammar configuration file format
- part of speech tagger
- Microsoft Windows version