LKB-FOS Update

John Carroll
University of Sussex, UK

DELPH-IN Summit, July 2022

Outline

Functionality: releases, new features, parameter changes

Chart mapping: algorithm, recipe

Comparative testing: processing steps, numbers of parses, diagnosis

Maintenance: internal improvements, bugs and fixes

Functionality

Three LKB-FOS releases since last year's summit

- 3 December 2021 (minor)
- 8 December 2021 (bug fix)
- 4 July 2022 (major)

Latest release includes chart mapping (token mapping and lexical filtering)

Parameter changes

- new parameter `*show-incomplete-lex-rule-chains*` controls whether parse chart window shows chains of lexical rule applications that are incomplete
- changed interpretation of parameter `*first-only-p*` – if zero, a complete parse/generate chart is created but no results are unpacked

Chart Mapping

Based on Adolphs *et al.* 2008 paper and DELPH-IN Summit 2009 slides

No published algorithm, so when in doubt I consulted ACE and PET source code, and examined trace of processing steps from PET

Implemented token mapping and lexical filtering; post-generation mapping will be added in future

Token mapping rule example:

```
suffix_apostrophe_tmr := suffix_punctuation_tmt &
[ +INPUT < [ +FORM ^(.*[^sS])$ ] >,
  +CONTEXT < [ +FORM "'" ] >,
  +OUTPUT < [ +FORM "${I1:+FORM:1}'" ] >,
  +POSITION "I1<C1, O1@I1, O1@C1" ].
```

Algorithm

The 'anchor' – the rightmost input/context argument – is computed for each rule

- each rule's anchor is moved across edges in the chart $L \rightarrow R$
- if anchor matches an edge, remaining args matched against other edges $R \rightarrow L$

When a rule fires

- if there are output edges, the rule is restarted with its anchor at the left vertex of the leftmost of these
- otherwise the rule is restarted at the current anchor position

Restarting a rule from scratch and at these vertex positions ensures that: (1) a rule cannot spuriously match edges it has just removed, and (2) a rule has the opportunity to match any new edges it has just added

Pseudocode and source code at
<http://svn.delph-in.net/lkb/branches/fos/src/main/chartmap.lsp>

Recipe

Set the relevant parameters listed in src/main/globals.lsp

```
;; token type:
(defparameter *token-type* 'token)

;; paths in token fs:
(defparameter *token-form-path* '(+FORM))
(defparameter *token-id-path* '(+ID))
(defparameter *token-from-path* '(+FROM))
(defparameter *token-to-path* '(+TO))
(defparameter *token-postags-path* '(+TNT +TAGS))
(defparameter *token-posprobs-path* '(+TNT +PRBS))

;; path to token feature structures in lexical items:
(defparameter *lexicon-tokens-path* '(TOKENS +LIST))
(defparameter *lexicon-last-token-path* '(TOKENS +LAST))

;; paths in chart mapping rules:
(defparameter *chart-mapping-context-path* '(+CONTEXT))
(defparameter *chart-mapping-input-path* '(+INPUT))
(defparameter *chart-mapping-output-path* '(+OUTPUT))
(defparameter *chart-mapping-position-path* '(+POSITION))
(defparameter *chart-mapping-jump-path* '(+JUMP))
```

Load generic LEs from a sub-lexicon called "gle"

```
(read-cached-sublex-if-available
   "gle" (lkb-pathname (parent-directory) "gle.tdl"))
```

Read in token mapping and lexical filtering rules

```
(loop for file in '(
      "tmr/gml.tdl" "tmr/ptb.tdl" "tmr/spelling.tdl" "tmr/ne1.tdl"
      "tmr/split.tdl" "tmr/ne2.tdl" "tmr/class.tdl" "tmr/ne3.tdl"
      "tmr/punctuation.tdl" "tmr/pos.tdl" "tmr/finis.tdl")
    do
    (read-token-mapping-file-aux (lkb-pathname (parent-directory) file)))
(read-lexical-filtering-file-aux (lkb-pathname (parent-directory) "lfr.tdl"))
```

Ensure `*parse-ignore-rules*` and `*repp-interactive*` are set appropriately

Script and settings files can use `#+:chart-mapping` and `#-:chart-mapping` to control whether to load chart mapping rules and set relevant variables. E.g.

```
#+:chart-mapping
(read-token-mapping-file-aux
  (lkb-pathname (parent-directory) "tmr/gml.tdl"))
```

To get a trace of processing steps, set the variable `*cm-debug*` to a non-nil value:

- `t` or `1` successful chart mapping rule applications only
- 2 additionally, regex match attempts
- 3 additionally, all attempts to match each rule

Comparative Testing

Cross-checking with PET

- compare traces of chart mapping steps

Cross-checking with ACE

- compare numbers of parses
- look for causes of missing / extra parses

Traces of chart mapping steps

Compared a few traces of chart mapping in LKB and PET – unsystematically and by hand

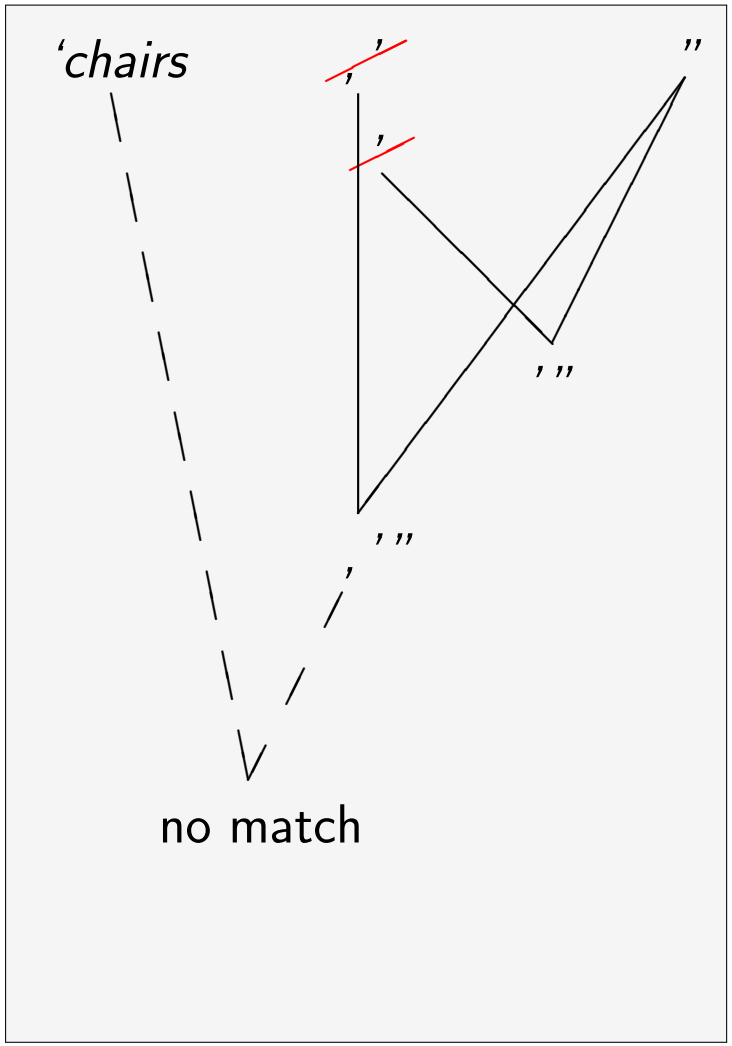
Found one discrepancy, in token mapping applied to *'chairs,'"* (as in the phrase *"those 'chairs,'" she said.*)

- `suffix_punctuation_tmr` deals with suffix punctuation, glueing trailing punctuation marks in turn to the end of the preceding token
- however, if the preceding token is itself a trailing punctuation mark, then a cluster of these characters can be formed which cannot then be attached to the preceding word ☹

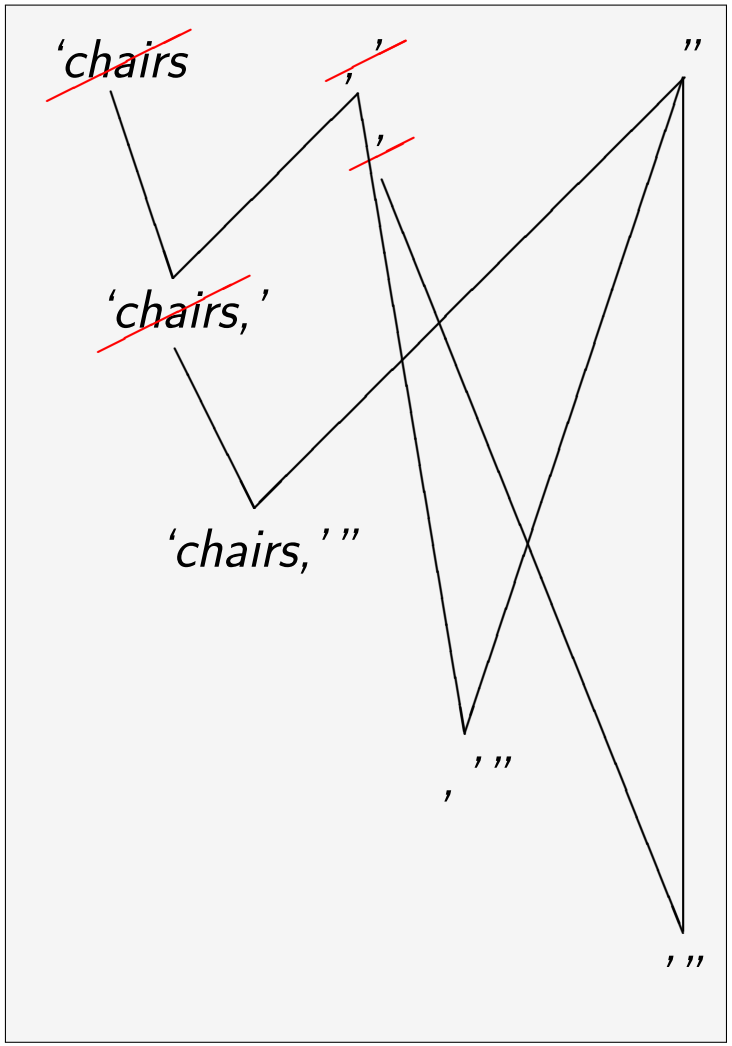
E.g. PET processes *'chairs,'"* R $\rightarrow$ L and gets stuck in a dead-end with the punctuation cluster *,'"* whereas LKB goes L $\rightarrow$ R and produces the intended result

Both strategies are consistent with the published descriptions of chart mapping

PET



LKB



Numbers of parses

Parsed Rondane test suite with LKB and ACE, specifying high resource limits and up to 100K best parses

Compared numbers of parses for items where neither system hit resource limit or returned as many as 100K parses

1103 such items; out of these, 24 had different numbers of parses

	#ACE	#LKB	Sentence
139	499	455	*The beauty of this place prompts us to pause often in order to soak it all in.*
257	9	7	*Experience Required*
382	18245	7834	*With the peak of Raudalstindane coming into view we will follow a chain of small lakes until we reach the larger lake of Raudalsvatnet.*
488	3	6	*To Bjørndalstinden*
509	45510	44892	*The second option was to get on the Melderskin - Bjørndalstindane ridge, but Petter believed a notch would cause more problems on that route.*
632	5	3	*Dinner included*
640	20	18	*Breakfast & dinner included* (also 648, 674, 696, 710, 719)
662	25	23	*Breakfast & lunch included*
683	7	5	*Breakfast included*
727	42981	38170	*"Alpine tundra" environments are encountered at an elevation of about 3,000 feet above sea level.*
728	32	26	*No hike on this trip exceeds an elevation of 4,500 above sea level.*
749	148	100	*At last I could see my final destination, the airport was clearly in view.*
924	37267	39037	*Weather forecast: cloudy with a cold wind blowing: wear jackets!*
977	74	44	*Once on top we can see a section of the route of this morning.*
1114	5041	4785	*Torger also pointed out the tiny summer farmhouse of this same family, a speck high on the precarious hillside.*
1138	16182	13338	*The summits are ~2000 m or more above sea level, and there are quite a number of smaller glaciers in the area.*
1145	22624	22600	*Some are on a self service basis, i.e. there are supplies for sale, but you have to cook yourself.*
1179	4442	3624	*It touches Treriksröset, where the borders of the three countries have one point in common.*
1191	1419	1379	*The swedish huts have a hutkeeper during season, and out of season one room is accessible.*

Missing / extra parses

Investigated differences in first half of test suite, and found 3 kinds of reason:

- *Experience Required* ACE 9 parses, LKB 7

  Two of the ACE parses (with `j_frg_c` top node) are filtered out in the LKB due to cyclic structure being created in unifiability check with roots (the *Dinner included* etc. items are similar)

- *To Bjørndalstinden* ACE 3 parses, LKB 6

  Both LKB and ACE have a lexical edge for *Bjørndalstinden* which undergoes `n_sg_ilr`; LKB has a further analysis of *Bjørndalstinden* which undergoes `n_pl-irreg-noaff_olr` due to an entry in the `irregs.tab` file

- *no hike exceeds 4,500 above sea level* ACE 10 parses, LKB 8

  In LKB, two of the ACE parses (those analysing *exceeds...level* with `hd-aj_int-unsl_c`) are filtered out by the idiom check

Maintenance

Data structures and algorithms

Faster access to agenda

Removed unnecessary indirection in parse chart edges

Reduced numbers of attribute comparisons in unification and subsumption check

- sort arcs in LEs and rules based on attributes' level of introduction
- combine arcs in unification results so they remain approximately ordered

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*
ACE	14:35	ACE emulated	10:21
LKB-FOS	24:32	LKB-FOS emulated	14:22
		LKB-FOS native	13:52

Development issues

McCLIM

- updating often throws up new issues, so binaries still at version of January 2022
- main outstanding bug: windows with a lot of content get garbled by scrolling
- have accumulated a set of patches to McCLIM to make it more friendly for LKB-FOS users

SBCL

- also stay a couple of releases behind bleeding edge
- worked around a performance degredation in unification caused by a change in how SBCL implements its store barrier

Reduced number of compiler warnings (also benefits Allegro CL), removed redundant files in svn

Bugs and fixes

DELPH-IN Discourse very useful for receiving bug reports and disseminating fixes – many thanks for clear reports and minimal reproducible examples

Also:

- found further cases where memory was not released promptly after errors and timeouts: FS 'unfilling', agenda handling
- fixed a few obscure, long-standing bugs: multi-word morphology vs. lexical rules, agenda item ordering

Summary in the README file

Summary

Development has continued over the past year

- chart mapping
- comparative testing
- maintenance and bug fixes

Still to do

- part of speech tagger?
- post-generation mapping rules
- selective unpacking 'grandparenting'
- unified grammar configuration file format
- Microsoft Windows version (perhaps without a GUI)