An application of AOC-posets: Indexing large corpuses for text generation under constraints

Alain Gutierrez Michel Chein
Marianne Huchard Pierre Pompidor

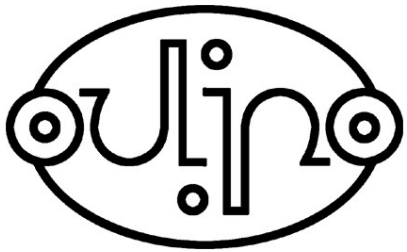
LIRMM, CNRS and Univ. Montpellier, France

ISMIS 2017, Warsaw, 2017, june 26-29

Outline

OULIPO

1960



Literary approach

Raymond Queneau, François Le Lionnais, Italo Calvino, Claude Berge, Georges Perec, Jacques Roubaud, Olivier Salon (....)

Experiment literary creation with formal constraints in writing, as a boost for imagination

Text generation under constraints

Lipogram: Exclude one or several letters of the alphabet

La disparition (G. Perec, a novel of $\sim$ 300 pages written without 'e')

Univocalic: Use a single vowel

No cool monsoons blow soft on Oxford dons, Orthodox, jog-trot, book-worm Solomons (C.C. Bombaugh, 1890)

N+i machine (J. Lescure): Replace each noun in a text with the *i-th* one following it in a dictionary

http://www.spoonbill.org/n+7/

N+0 I am a cat and I see a mouse
N+1 I am a cataclysm and I see a mousetrap
N+2 I am a catacomb and I see a moustache
N+3 I am a catalogue and I see a mouth
N+4 I am a catalyst and I see a mouthful
N+5 I am a catamaran and I see a mouthpiece
N+6 I am a catapult and I see a mouthwash
N+7 I am a cataract and I see a movement

ALAMO

1981 Alamo

Atelier de littérature assistée par les mathématiques et les ordinateurs
Workshop of Literature Assisted by Mathematics and Computers
Created by Paul Braffort and Jacques Roubaud
Current president Guy Chaty

http://www.alamo.free.fr/

Lapal

CogiText = *CoGui + Text*
CoGui, visual tool for building knowledge bases
http://www.lirmm.fr/cogui/

Illustration with a constrained text parody

The fox and the crow
J. de la Fontaine, 1668

Original	**Original**
Le corbeau et le renard. Maître Corbeau, sur un arbre perché, Tenait en son bec un fromage.	The crow and the fox. Mister Crow, perched on a tree, was holding in his beak a cheese.
Parody	**Parody**
Le barbot et le fouinard. Maitre barbot, sur un marbre torché, Tenait en son bec un dommage	The woe and the vox. Mister woe, lurched on a knee, was holding in its creek a sneeze.

Outline

Corpuses

DELA

102 073 lemmas, 683 824 inflected forms

http://infolingu.univ-mlv.fr/

Morphalou

8790 verb lemmas

http://www.cnrtl.fr/lexiques/morphalou/LMF-Morphalou.php

Phonetic syllabification (handmade)

"accéléraient" $\rightarrow$ `ak se le rE`

641 phonetic rules (built from lexicon Descartes analysis)

1399 lemmas with exceptional phonetics

Corpuses

Corpus element

(txt,"home")	(rhyme2,"oM")	(gender,"_neutral")	(nbsyl, 3)

Corpus schema

(txt,string)	(rhyme2,string)	(gender,string)	(nbsyl,integer)

Corpus mappings

`reverse`	$e.txt = "crow" \rightarrow$ `e.reverse = "worc"`
`nth(i)`	$e.txt = "cat" \rightarrow$ `e.nth(6)= "catapult"`
	(.....)

Outline

Production scheme

Production template (french)	**Production template (transposed)**
Le $\{X_1.txt\}$ et le $\{X_2.txt\}$. Maître $\{X_1.txt\}$, sur un $\{X_3.txt\}$ $\{X_4.txt\}$ Tenait en son $\{X_5.txt\}$ un $\{X_6.txt\}$	The $\{X_1.txt\}$ and the $\{X_2.txt\}$. Mister $\{X_1.txt\}$, $\{X_4.txt\}$ on a $\{X_3.txt\}$, was holding in its $\{X_5.txt\}$ a $\{X_6.txt\}$.
Constraint set (french)	**Constraint set (transposed)**
X_1=element(corpusNoun) X_1.rhyme3="Rbo" X_1.nbsyl=2 X_1.gender="_masculine" X_1.number="_singular" X_2=element(corpusNoun) X_2.rhyme3="naR" X_2.nbsyl=3 .... X_3.gender=X_4.gender X_3.number=X_4.number .... X_5=element(corpusNoun) .... X_6=element(corpusNoun)	X_1=element(corpusNoun) X_1.rhyme2=**"oW"** X_1.nbsyl=**1** X_1.gender=**"_neutral"** X_1.number="_singular" X_2=element(corpusNoun) X_2.rhyme1=**"oX"** X_2.nbsyl=**1** .... X_3.gender=X_4.gender X_3.number=X_4.number .... X_5=element(corpusNoun) .... X_6=element(corpusNoun)

Outline

Steps

1. Build AOC-poset
   - index to corpuses
   - offline
2. Assign values to variables attributes (key-value pairs)
   - e.g. assign value to $X_i.gender$
3. Assign terms (corpus elements) to variables
   - e.g. assign term to X_i

Formal context

Offset (text) × key-value	gender _masculine	number _singular	nbsyl 1	nbsyl 2	nbsyl 3	rhyme3 naR	rhyme3 Rbo	… …
164555 (renard)	x	x		x		x		…
348 (fouinard)	x	x		x		x		…
110976 (corbeau)	x	x		x			x	…
345724 (barbot)	x	x		x			x	…
734657 (turbot)	x	x		x			x	…
12456 (arbre)	x	x		x				…
78347 (marbre)	x	x		x				…
1110723 (bec)	x	x	x					…
34677 (fromage)	x	x			x			…
125044 (dommage)	x	x			x			…
.....	…	…	…	…	…	…	…	…

Formal concept

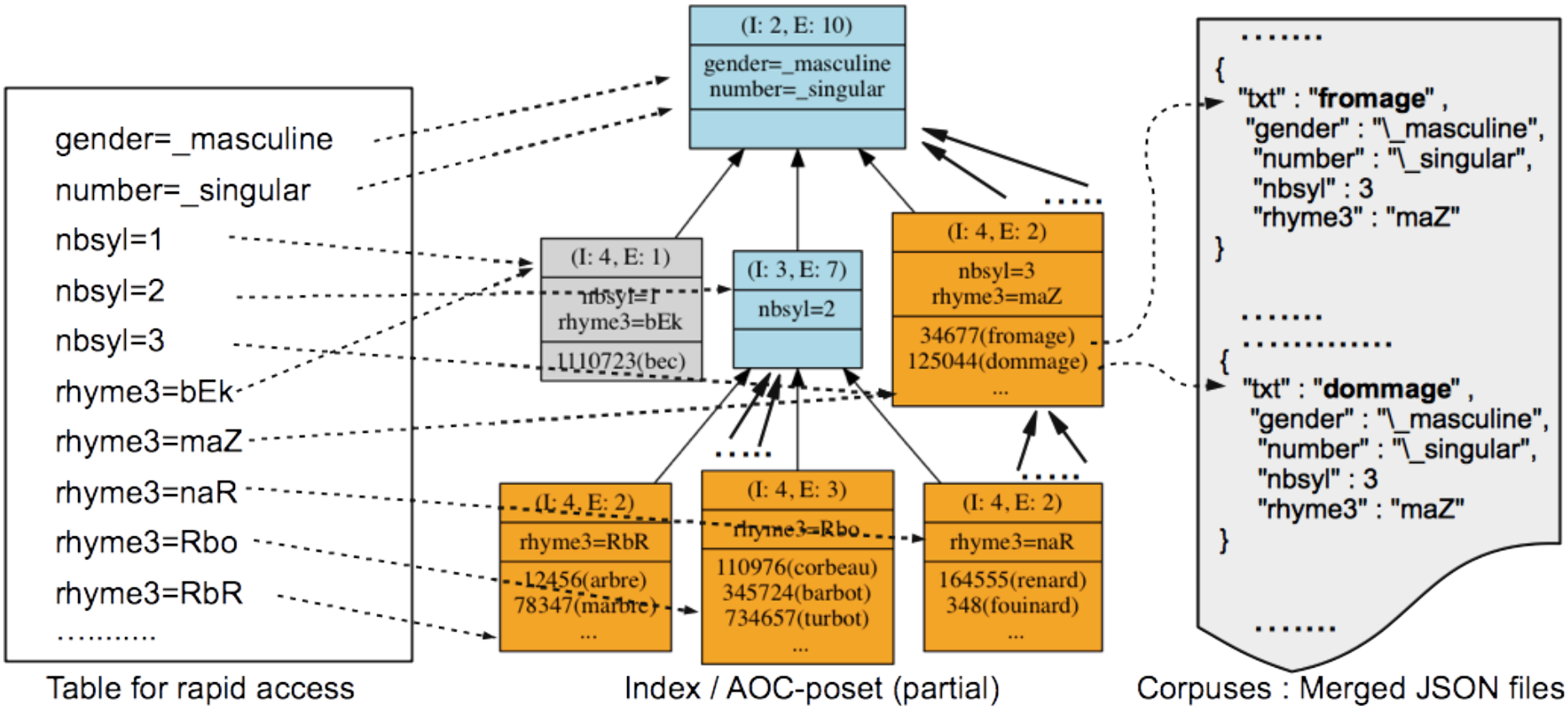
Offset (text) × key-value	gender _masculine	number _singular	nbsyl 1	nbsyl 2	nbsyl 3	rhyme3 naR	rhyme3 Rbo	… …
164555 (renard)	x	x		x		x		…
348 (fouinard)	x	x		x		x		…
110976 (corbeau)	x	x		x			x	…
345724 (barbot)	x	x		x			x	…
734657 (turbot)	x	x		x			x	…
12456 (arbre)	x	x		x				…
78347 (marbre)	x	x		x				…
1110723 (bec)	x	x	x					…
34677 (fromage)	x	x			x			…
125044 (dommage)	x	x			x			…
.....	…	…	…	…	…	…	…	…

Concept specialization

Offset (text) × key-value	gender _masculine	number _singular	nbsyl 1	nbsyl 2	nbsyl 3	rhyme3 naR	rhyme3 Rbo	… …
164555 (renard)	x	x		x		x		…
348 (fouinard)	x	x		x		x		…
110976 (corbeau)	x	x		x			x	…
345724 (barbot)	x	x		x			x	…
734657 (turbot)	x	x		x			x	…
12456 (arbre)	x	x		x				…
78347 (marbre)	x	x		x				…
1110723 (bec)	x	x	x					…
34677 (fromage)	x	x			x			…
125044 (dommage)	x	x			x			…
.....	…	…	…	…	…	…	…	…

Offset (text) × key-value	gender _masculine	number _singular	nbsyl 1	nbsyl 2	nbsyl 3	rhyme3 naR	rhyme3 Rbo	… …
164555 (renard)	x	x		x		x		…
348 (fouinard)	x	x		x		x		…
110976 (corbeau)	x	x		x			x	…
345724 (barbot)	x	x		x			x	…
734657 (turbot)	x	x		x			x	…
12456 (arbre)	x	x		x				…
78347 (marbre)	x	x		x				…
1110723 (bec)	x	x	x					…
34677 (fromage)	x	x			x			…
125044 (dommage)	x	x			x			…
.....	…	…	…	…	…	…	…	…

AOC-poset (concepts introducing an Attribute or an Object)



Computation of key-value pairs for corpus variables

Group by equality

Case 1 - The group contains a value

Constraints	**Group**
X_1.gender=X_2.gender X_1.gender="_masculine"	X_1.gender X_2.gender "_masculine"

Computation of key-value pairs for corpus variables

Group by equality Case 2 - The group does not contain a value

Constraints	Group
X_3.nbsyl = X_4.nbvowels	X_3.nbsyl X_4.nbvowels

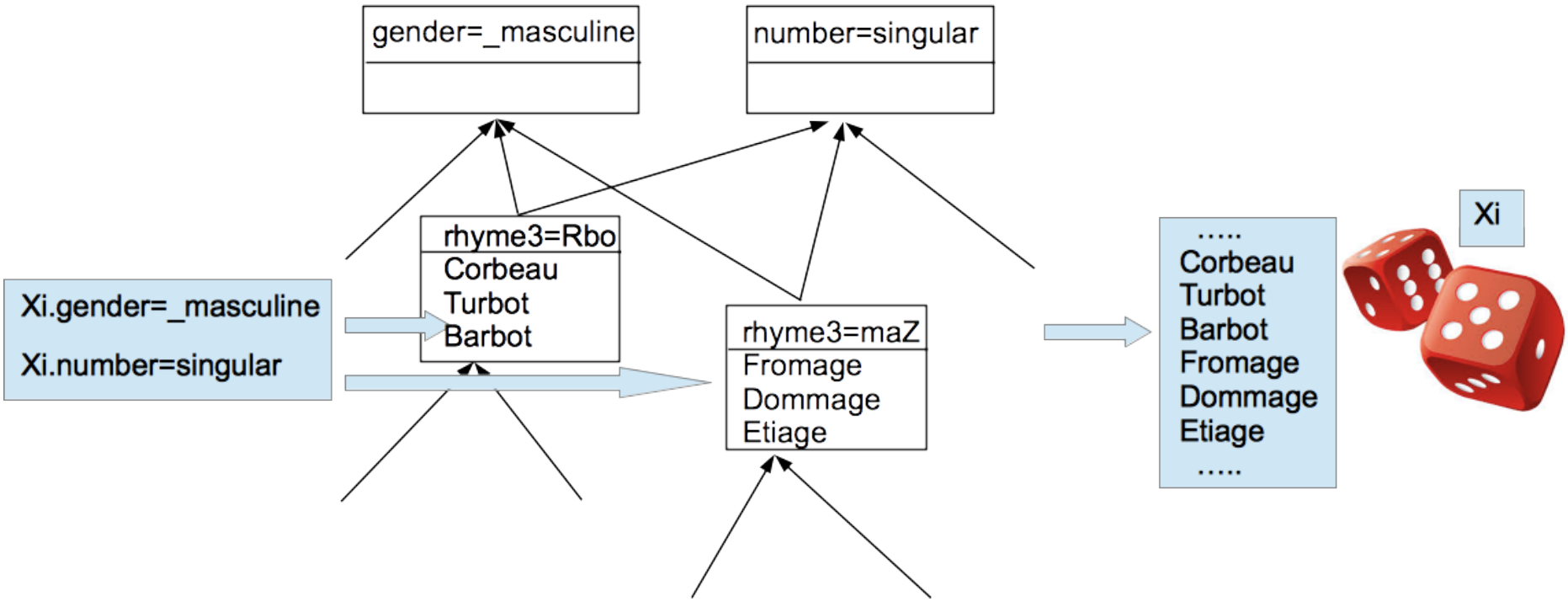
Variable	Corpus	Possible values
X_3.nbsyl	DELA nouns	$\{1, 2, ...14\}$
X_4.nbvowels	DELA adjective	$\{1, 2, ...6\}$

Intersection

$\{1, 2, ...14\} \cap \{1, 2, ...6\} = \{1, 2, ...6\}$

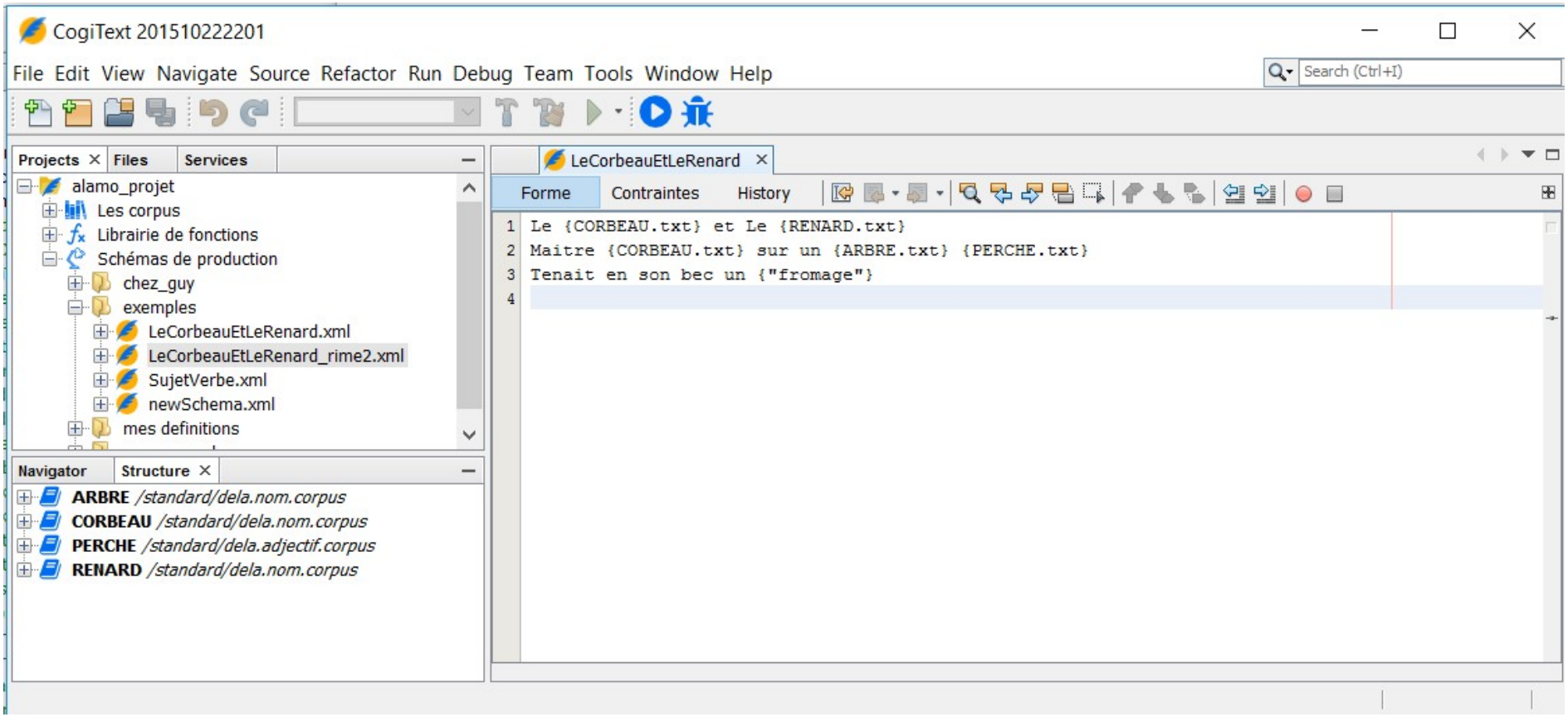
Weighted random sampling among the values v_i of intersection	Weights = # pairs (noun, adj) with v_i syllables and v_i vowels
1	n_1
2	n_2
....	....
6	n_6

Assignment of corpus elements to variables

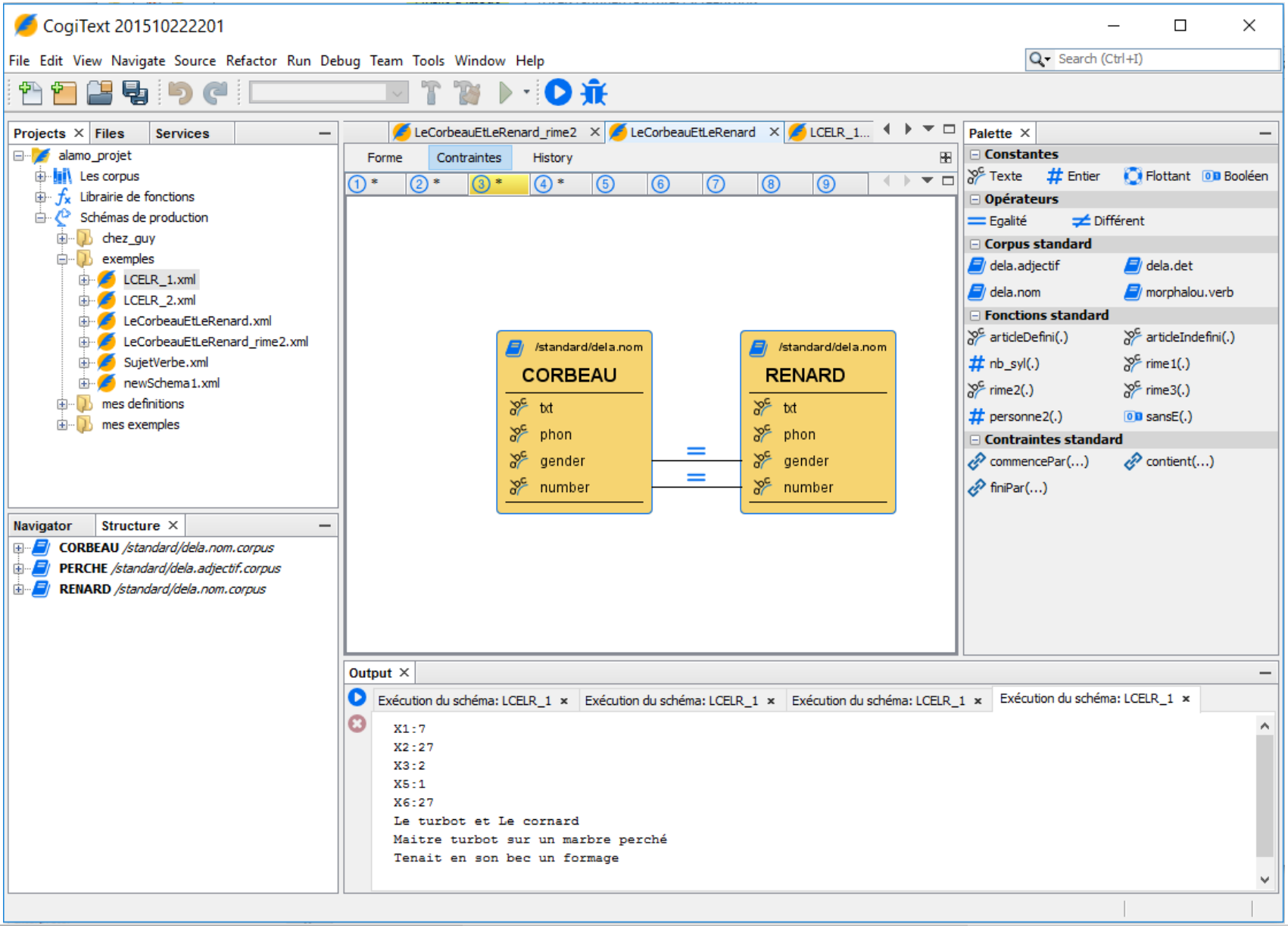


Outline

Implementation



Implementation



Experiments

key-value pairs for **rhyme3+nbsyl+gender+number with** filtering

#elements	#key-value pairs	density	building matrix ex. time	#concepts
137276	10	0.17	50s	56
Time	Ceres (ms)	Pluton (ms)	Hermes (ms)	Ares (ms)
BitSet	1229	2057	3124	26445
HashSet	1327	425	85887 $\sim$ 1,5min	36186

key-value pairs for **rhyme3+nbsyl+gender+number without** filtering

#elements	#key-value pairs	density	building matrix ex. time	#concepts
160268	4800	8.32E-4	50s	33669
Time	Ceres (ms)	Pluton (ms)	Hermes (ms)	Ares (ms)
BitSet	216152	1884040	1422808	4018082 $\sim$ 1h
HashSet	138069 $\sim$ 2min	400936	580275	3635452

Time for a text production

- For a 3-phonemes search
  #poss.: X1:7, X2:27, X3:2, X4:3, X5:1, X6:27
  - filtered data: 787 ms
    including a 3 ms traversal
  - non filtered data: 1571 ms
    including a 37 ms traversal

Outline

Conclusion/Perspectives

Conclusion

- Assist constrained literary text generation with:
  - corpuses, corpus schemas
  - production patterns, constraints
  - AOC-poset indexing structure

Perspectives

- Understanding algorithm/data structure practical complexity
- Deal with more complex constraints
- Enrich the mapping library
- On-the-fly AOC-poset generation for specific production schemes

Any questions?

Any questioners?
Any questionnaires?
Any queues?
Any quibbles?
Any quiches?
Any quickies?
Any quicksands?
Any quids?
Any quiets?
Any quills?
Any quilts?
Any quins?
Any quinces?
Any quintets?
Any quintuplets?

Acknowledgements
Guy Chaty