

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

- 1 Context
- 2 Corporuses
- 3 Production scheme
- 4 Indexing and selecting
- 5 Experiments
- 6 Conclusion

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 ~ 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

- 1 Context
- 2 Corpu**s****
- 3 Production scheme
- 4 Indexing and selecting
- 5 Experiments
- 6 Conclusion

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" → 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

- 1 Context
- 2 Corpora
- 3 Production scheme**
- 4 Indexing and selecting
- 5 Experiments
- 6 Conclusion

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) 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)	Constraint set (transposed) 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

- 1 Context
- 2 Corporuses
- 3 Production scheme
- 4 Indexing and selecting**
- 5 Experiments
- 6 Conclusion

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			...
.....

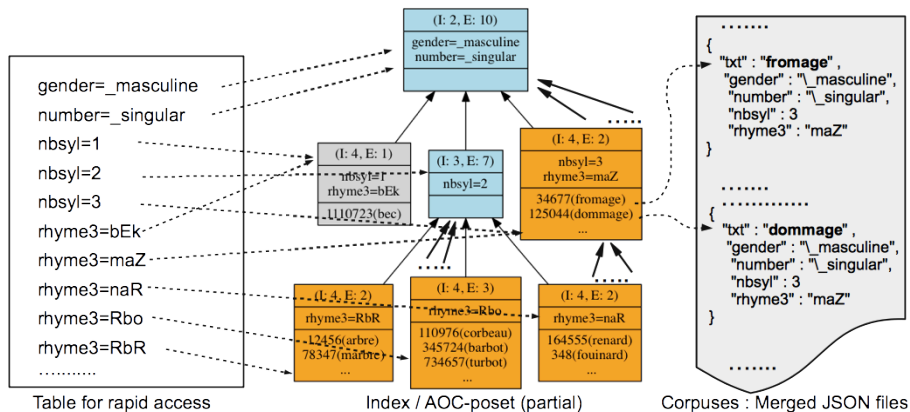
Formal concept

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	<u>gender</u> masculine	<u>number</u> singular	nbsyl 1	<u>nbsyl</u> 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	<u>gender</u> masculine	<u>number</u> singular	nbsyl 1	<u>nbsyl</u> 2	nbsyl 3	rhyme3 naR	<u>rhyme3</u> 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
$X_1.\text{gender}=X_2.\text{gender}$ $X_1.\text{gender}=\text{"_masculine"}$	$X_1.\text{gender}$ $X_2.\text{gender}$ "_masculine"

Computation of key-value pairs for corpus variables

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

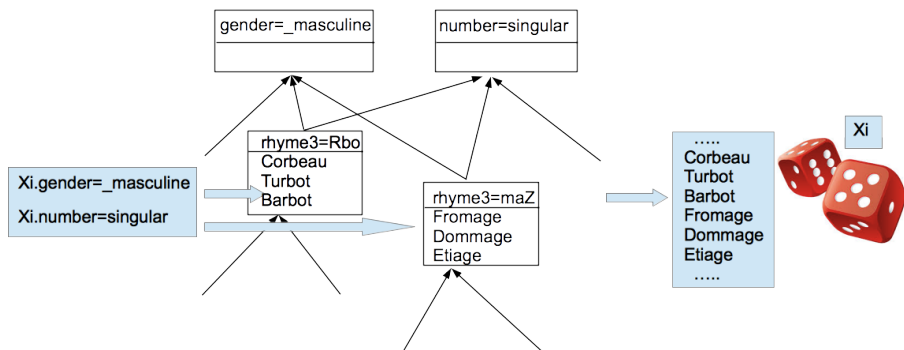
Constraints	Group
$X_3.nbsyl = X_4.nbvowels$	$X_3.nbsyl$ $X_4.nbvowels$

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

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

- 1 Context
- 2 Corporuses
- 3 Production scheme
- 4 Indexing and selecting
- 5 Experiments**
- 6 Conclusion

Implementation

The screenshot shows the CogiText IDE interface. The title bar reads "CogiText 201510222201". The menu bar includes "File", "Edit", "View", "Navigate", "Source", "Refactor", "Run", "Debug", "Team", "Tools", "Window", and "Help". A search bar on the right contains "Search (Ctrl+F)".

The left sidebar shows a project tree for "alamo_projet" with folders like "Les corpus", "Librairie de fonctions", and "exemples". Under "exemples", the file "LeCorbeauEtLeRenard.xml" is selected.

The main editor window, titled "LeCorbeauEtLeRenard", shows the following text:

```

1 Le {CORBEAU.txt} et Le {RENARD.txt}
2 Maitre {CORBEAU.txt} sur un {ARBRE.txt} {PERCHE.txt}
3 Tenait en son bec un {"fromage"}
4

```

Below the editor, the "Navigator" pane shows a structure view with the following entries:

- ARBRE /standard/dela.nom.corpus
- CORBEAU /standard/dela.nom.corpus
- PERCHE /standard/dela.adjectif.corpus
- RENARD /standard/dela.nom.corpus

Implementation

Cogitext 20151022201

File Edit View Navigate Source Refactor Run Debug Team Tools Window Help

Search (Ctrl+I)

Projects Files Services

- alano_projet
 - Les corpus
 - Librairie de fonctions
 - Schémas de production
 - chez_guy
 - exemples
 - LCELR_1.xml
 - LCELR_2.xml
 - LeCorbeauEtLeRenard.xml
 - LeCorbeauEtLeRenard_rime2.xml
 - SujetVerbe.xml
 - newSchema1.xml
 - mes definitions
 - mes exemples

Navigator Structure

- CORBEAU /standard/dela.nom.corpus
- PERCHE /standard/dela.adjectif.corpus
- RENIARD /standard/dela.nom.corpus

Forme Contraintes History

1 2 3 4 5 6 7 8 9

Diagram showing two classes: CORBEAU and RENARD, both inheriting from /standard/dela.nom. Both classes have attributes: bt, phon, gender, and number.

Palette

- Constantes
 - Texte # Entier Flottant Boolean
- Opérateurs
 - Egalité Différent
- Corpus standard
 - dela.adjectif dela.det
 - dela.nom morphalou.verb
- Fonctions standard
 - articleDefini(.) articleIndefini(.)
 - nb_syl(.) rime1(.)
 - rime2(.) rime3(.)
 - personne2(.) sansE(.)
- Contraintes standard
 - commencePar(...) content(...)
 - finPar(...)

Output

Exécution du schéma: LCELR_1 x Exécution du schéma: LCELR_1 x Exécution du schéma: LCELR_1 x Exécution du schéma: LCELR_1 x

```

X1:7
X2:27
X3:2
X5:1
X6:27
Le turbot et Le cornard
Maitre turbot sur un marbre perché
Tenait en son bec un fromage
  
```

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
<i>Time</i>	Ceres (ms)	Pluton (ms)	Hermes (ms)	Ares (ms)
BitSet	1229	2057	3124	26445
HashSet	1327	425	85887 ~ 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
<i>Time</i>	Ceres (ms)	Pluton (ms)	Hermes (ms)	Ares (ms)
BitSet	216152	1884040	1422808	4018082 ~ 1h
HashSet	138069 ~ 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

- 1 Context
- 2 Corporuses
- 3 Production scheme
- 4 Indexing and selecting
- 5 Experiments
- 6 Conclusion**

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