Plate-formes logicielles pour le TAL 1 : expressions régulières et recherche de motifs dans Unitex

Karën Fort

karen.fort@paris-sorbonne.fr / http://karenfort.org

18 novembre 2016

Quelques sources d'inspiration

- Manuel d'`Unitex` :
  http://www-igm.univ-mlv.fr/~unitex/index.php?page=4
- Cours de M. Constant, Université de Marne-la-Vallée

Exemples de recherches de motifs

- un mot (*juger*) ou séquence de mots (*pomme de terre*)
- toutes les formes fléchies associées à une forme de base (<*juger*> = juge, juges, jugeons, ...)
- formes appartenant à une catégorie grammaticale avec informations flexionnelles : <*N*>, <*N :ms*>, <*V :K*>
- motifs complexes : <*DET :ms*> <*N :ms*>
- expressions régulières : *je+tu+il+elle+on+nous+vous+ils+elles*
- automates sous la forme de graphes

À quoi ça sert ?

- filtrage et routage de documents
- extraction d'information
- aide à la traduction

Lancer Unitex (rappel)

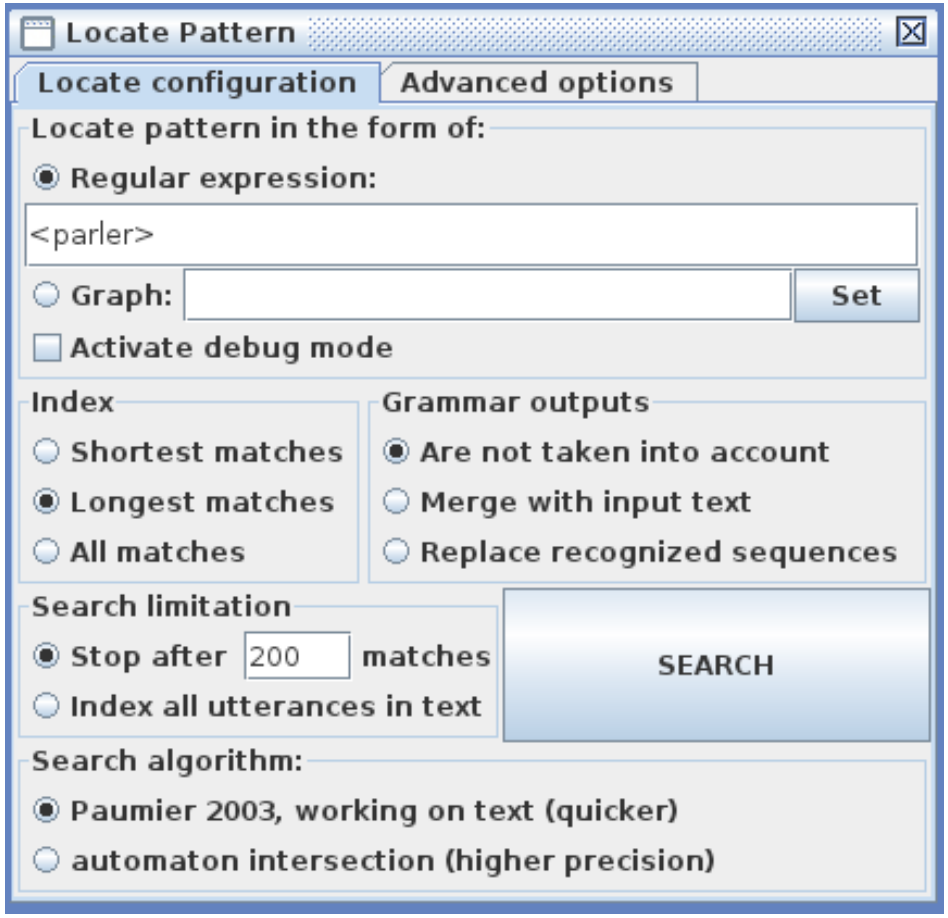
- lancez Unitex
- vérifiez que la liste de langues est correcte
- vérifiez que vous avez un répertoire personnel (Info/Preferences/Directories), qui n'est pas celui de l'installation
- modifiez l'encodage par défaut (Info/Preferences/Encoding) et mettre UTF8
- ouvrez le *Tour du monde en 80 jours* (TDM) avec le prétraitement par défaut

Recherches simples

Parole, parole, parole

1. rechercher le motif *parler* en cliquant sur `Locate Pattern` dans le menu `Text`
   - regarder le résultat avec le concordancier
   - modifier les différentes options et observer les résultats
2. même question avec le motif *<parler>*
3. même question avec le motif *<V :P3p>*
4. à quoi correspondent les motifs précédents ?

Locate Pattern



Expressions *régulières* ou *rationnelles*

Une expression rationnelle peut être :

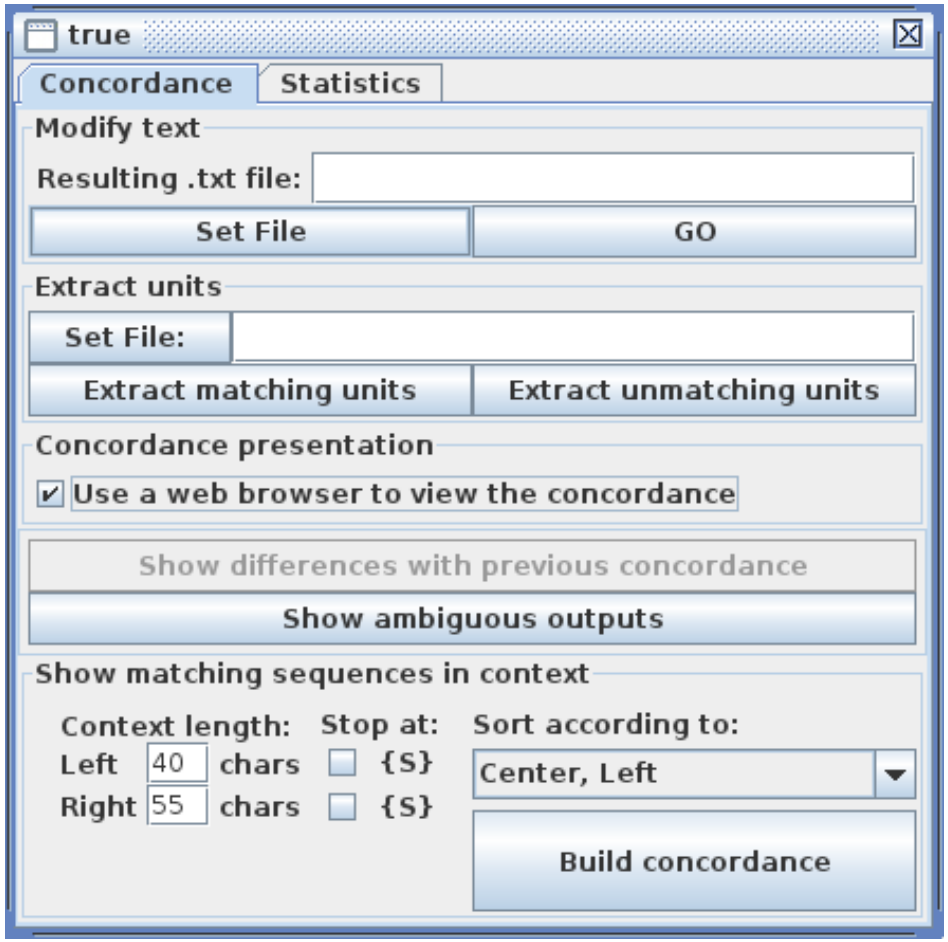
- une unité lexicale (*livre*) ou un masque lexical (<*manger.V*>)
- une position particulière du texte : le début (^) ou la fin (*$*)
- la concaténation de deux expressions rationnelles (*je mange*)
- l'union de deux expressions rationnelles (*Pierre+Paul*)
- l'étoile de Kleene d'une expression rationnelle (*très**)

Concordance

Définition

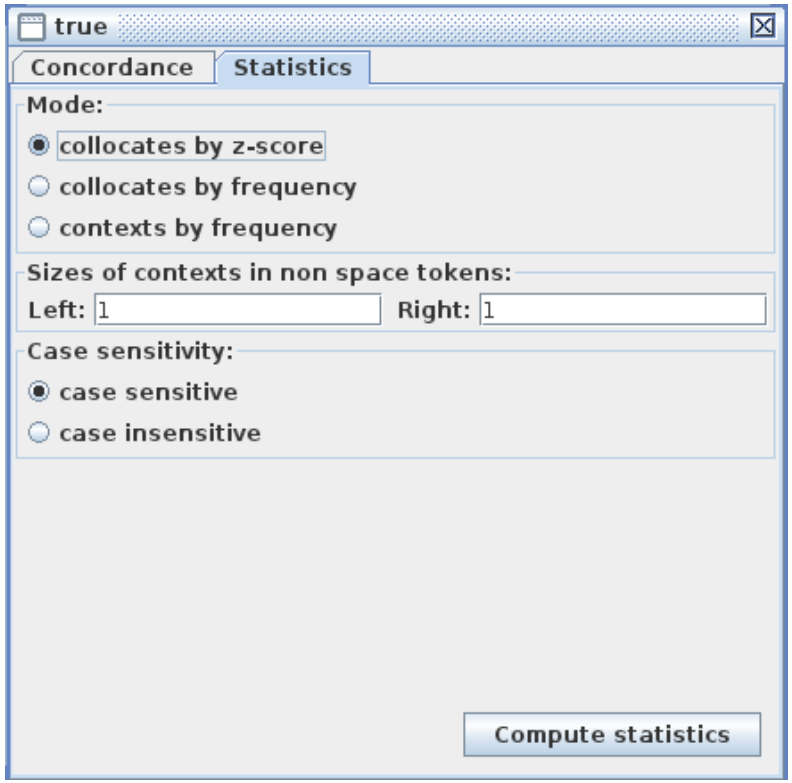
≪ Répertoire des exemples rencontrés pour chaque mot et donnant pour chaque occurrence un contexte de trois lignes, le mot étudié figurant obligatoirement dans la ligne du milieu ≫
http://www.cnrtl.fr/definition/concordance (TLFi)

Lancer le concordancier



Concordance sur le motif <*parler*>

Statistiques de collocations



Collocations du motif <*parler*>

Statistics

Collocate	Occurrences in corpus	Occurrence in match context	z-score
LUI	2	1	20.841
Ne	5	1	13.136
UN	8	1	10.349
entendu	8	1	10.349
demain	12	1	8.411
me	52	2	7.951
PASSEPARTOUT	14	1	7.769
pendant	62	2	7.239
pas	457	5	6.201
faire	85	2	6.099
pouvait	86	2	6.06
avec	208	3	5.663
Elle	31	1	5.118
toujours	42	1	4.34
ainsi	44	1	4.23
lui	342	3	4.174
avait	372	3	3.95
garçon	54	1	3.772
aussi	67	1	3.333
vous	269	2	3.052
ne	554	3	2.977
encore	99	1	2.634
même	133	1	2.173
ce	452	2	2.064
sans	159	1	1.918
on	201	1	1.605
d	726	2	1.286
Le	270	1	1.243
Fix	274	1	1.226
en	784	2	1.167
Il	308	1	1.091
de	2807	5	1.01
à	1695	3	0.766
qui	572	1	0.427
se	580	1	0.413
!	618	1	0.348
un	824	1	0.058
.	4566	5	-0.104
l	1137	1	-0.267
le	1608	1	-0.626

Opérateurs

- concaténation :
  - point : *<DET>.<N>* (reconnaît un déterminant suivi par un nom)
  - espace : *le <A> chat* (reconnaît l'unité lexicale *le*, suivie d'un adjectif et de l'unité lexicale *chat*)
  - NOTE sur les parenthèses : servent de délimiteurs
- union :
  - + : *chat+chien <v>* (reconnaît l'unité lexicale *chat* ou *chien*, suivie par un verbe)
  - NOTE sur epsilon : *le (petit+<E>) chat* (reconnaît les séquences *le chat* et *le petit chat*)
- étoile de Kleene : * (permet de reconnaître zéro, une ou plusieurs occurrences d'une expression)
  - *il fait très* froid* : reconnaît *il fait froid*, *il fait très froid*, *il fait très très froid*, etc
  - prioritaire sur les autres opérateurs
  - parenthèses pour appliquer l'étoile à une expression complexe

Recherches de motifs complexes

Rechercher dans le TDM

- toutes les occurrences des pronoms personnels *(je, tu, il, . . .)*
- toutes les occurrences des pronoms personnels qui sont suivis par un verbe
- toutes les suites d'au moins 3 adjectifs (A) ; qu'observez-vous de surprenant ?
- toutes les suites de noms. Pourquoi le motif $<N>^*$ produit-il une erreur ? Que faire pour l'éviter ?

Codes grammaticaux usuels

code	Sigification	exemple
A	adjectif	fabuleux
ADV	adverbe	réellement, à la longue
CONJC	conjonction de coordination	mais
CONJS	conjonction de subordination	puisque, à moins que
DET	déterminant	ses, trente-six
INTJ	interjection	adieu, mille millions de mille sabords
N	nom	prairie, vie sociale
PREP	préposition	sans, à la lumière de
PRO	pronom	tu, elle-même
V	verbe	continuer, copier-coller

Codes flexionnels usuels

Code	Signification
m	masculin
f	féminin
n	neutre
s	singulier
p	pluriel
1, 2, 3	1 ère, 2 ème, 3 ème personne
P	présent de l'indicatif
I	imparfait de l'indicatif
S	présent du subjonctif
T	imparfait du subjonctif
Y	présent de l'impératif
C	présent du conditionnel
J	passé simple
W	infinitif
G	participe présent
K	participe passé
F	futur

Codes sémantiques usuels

Code	Signification	Exemple
z1	langage courant	blague
z2	langage spécialisé	sépulcre
z3	langage très spécialisé	houer
Abst	abstrait	bon goût
Anl	animal	cheval de race
AnlColl	animal collectif	troupeau
Conc	concret	abbaye
ConcColl	concret collectif	décombres
Hum	humain	diplomate
HumColl	humain collectif	vieille garde
t	verbe transitif	foudroyer
i	verbe intransitif	fraterniser
en	particule pré-verbale (PPV) obligatoire	en imposer
se	verbe pronominal	se marier
ne	verbe à négation obligatoire	ne pas cesser de

Utiliser les informations grammaticales|flexionnelles|sémantiques

- les codes grammaticaux sont écrits en majuscules et entre <>
- les informations . . . sont précédées de . . .
  - sémantiques : « + »
  - flexionnelles : « : »

Attention

Les codes grammaticaux et sémantiques précèdent les codes flexionnels

Recherches utilisant les informations grammaticales|flexionnelles|sémantiques

Rechercher dans le TDM

- tous les adjectifs au féminin pluriel
- tous les noms possédant le trait sémantique ≪ humain collectif ≫
- tous les verbes à l'imparfait du langage courant

Recherches complexes utilisant la concaténation et l'union

Rechercher dans le TDM

- tous les verbes, soit à l'imparfait, soit au présent ou à l'imparfait du subjonctif

Méta motifs `Unitex`

- $< E >$: mot vide, ou epsilon. Reconnaît la séquence vide
- $< TOKEN >$: n'importe quelle unité lexicale sauf l'espace
- $< MOT >$: n'importe quelle unité lexicale formée de lettres
- $< MIN >$: [...] de lettres minuscules
- $< MAJ >$: [...] de lettres majuscules
- $< PRE >$: [...] de lettres et commençant par une majuscule
- $< DIC >$: n'importe quel mot figurant dans les dictionnaires du texte
- $< SDIC >$: [...] mot simple [...]
- $< CDIC >$: [...] mot composé [...]
- $< NB >$: n'importe quelle suite de chiffres contigus

Négation et interdiction

- ! (immédiatement après <) : négation d'un motif, possible sur :
  - les métas <MOT>, <MIN>, <MAJ>, <PRE>, <DIC>
  - les masques lexicaux ne comportant que des codes grammaticaux, sémantiques ou flexionnels ($<!V + z3 : P3 >$)
- $\sim$: exclut des codes ($<A{\sim}z3>$ reconnaît toutes les entrées qui ont le code *A* sans le code *z3*)
- # : interdit la présence de l'espace

Recherches utilisant les négations

ou pas

Rechercher dans le TDM

- tous les mots qui ne sont pas dans le dictionnaire
- tous les mots qui ne sont pas écrits tout en minuscules
- tous les noms non humains

Recherches à l'aide de méta motifs

ou pas

Rechercher

- tout les mots commençant par une majuscule
- tous les mots qui possèdent le trait sémantique ≪ concret ≫

Filtres morphologiques Unitex

Format

motif <<motif morphologique>>
sous la forme d'expressions régulières au format POSIX

(voir http://fr.wikipedia.org/wiki/Expression_rationnelle#Expressions_rationnelles_.C3.A9tendues_POSIX)

Par défaut, un filtre morphologique tout seul s'applique au méta <TOKEN>, c'est-à-dire à n'importe quelle unité lexicale sauf l'espace.

Filtres simples

- $<< ss >>$: contient ss
- $<< \hat{}a >>$: commence par a
- $<< ez\$ >>$: finit par ez
- $<< a.s >>$: contient a suivi par un caractère quelconque, suivi par s
- $<< a.^*s >>$: contient a suivi par un nombre de caractères quelconque, suivi par s
- $<< ss|tt >>$: contient ss ou tt
- $<< [aeiouy] >>$: contient une voyelle non accentuée
- $<< [aeiouy]3,5 >>$: contient une séquence de voyelles non accentuées, de longueur comprise entre 3 et 5
- $<< es? >>$: contient e suivi par un s facultatif
- $<< ss[\hat{}e]? >>$: contient ss suivi par un caractère qui n'est pas une voyelle e

Filtres plus complexes

- <<[ai]ble$>> : finit par *able* ou *ible*
- <<^([rst][aeiouy]){2,}$>> : mot formé de 2 ou plus séquences commençant par un *r*, *s* ou *t* suivi d'une voyelle non accentuée

Filtres plus complexes

Lorsqu'un filtre suit immédiatement un motif, il s'applique à ce qui est reconnu par le motif :

- <V :K><<i$>> : participe passé finissant par *i*
- <CDIC><< .* >> : mot composé contenant deux espaces
- <A :fs><<^pro>> : adjectif féminin singulier commençant par *pro*

Recherches utilisant les filtres

Rechercher dans le TDM

- tous les mots qui commencent par *anti* ou *pro*, suivis par un tiret facultatif
- tous les mots composés contenant un tiret
- tous les mots qui ne sont pas dans le dictionnaire et qui se terminent par *es*

Savoir :

- faire des recherches :
  - en utilisant les informations fournies par les dictionnaires (code grammatical, flexion, sémantique)
  - en utilisant la négation
  - en utilisant les méta motifs
  - en utilisant la concaténation et l'union
  - des filtres morphologiques
- et en visualiser le résultat à l'aide du concordancier

Recherches avancées

Exercice

Écrivez des expressions régulières permettant de rechercher :

- tous les adjectifs qui ne sont pas très spécialisés
- tous les verbes un peu ou très spécialisés, soit au participe passé, soit à l'infinitif
- toutes les séquences
  - commençant par le verbe avoir (et)
  - se terminant par un verbe au participe passé (et)
  - dans lesquelles peuvent s'insérer des séquences quelconques de mots entre virgules (*eût, au contraire, perdu*)
- tous les verbes au subjonctif passé ou présent, contenant *uiss*