



Unitex : expressions régulières et recherche de motifs

Karën Fort

karen.fort@sorbonne-universite.fr / <https://members.loria.fr/KFort/>



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

Sources

Introduction

Recherche de motifs ?

Rappel

Se jeter à l'eau

Concordancier

Recherche de motifs par expressions régulières

Recherche de motifs par filtres morphologiques

Pour finir

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 ($\langle \textit{juger} \rangle = \textit{juge, juges, jugeons, ...}$)
- ▶ formes appartenant à une catégorie grammaticale avec informations flexionnelles : $\langle N \rangle$, $\langle N :ms \rangle$, $\langle V :K \rangle$
- ▶ motifs complexes : $\langle DET :ms \rangle \langle N :ms \rangle$
- ▶ 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

The image shows a software dialog box titled "Locate Pattern" with a close button in the top right corner. The dialog is divided into two tabs: "Locate configuration" (selected) and "Advanced options".

Under "Locate configuration", the text "Locate pattern in the form of:" is followed by three radio button options:

- Regular expression: Below this is a text input field containing the string "<parler>".
- Graph: Next to this is an empty text input field and a "Set" button.
- Activate debug mode

Below these are two columns of radio button options:

- Index**
 - Shortest matches
 - Longest matches
 - All matches
- Grammar outputs**
 - Are not taken into account
 - Merge with input text
 - Replace recognized sequences

Next is the "Search limitation" section with two radio button options:

- Stop after matches
- Index all utterances in text

To the right of these options is a large blue button labeled "SEARCH".

At the bottom is the "Search algorithm:" section with two radio button options:

- Paumier 2003, working on text (quicker)
- automaton intersection (higher precision)

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**)

Sources

Introduction

Concordancier

Présentation

Application

Statistiques

Recherche de motifs par expressions régulières

Recherche de motifs par filtres morphologiques

Pour finir

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

The image shows a software window titled 'true' with a close button in the top right corner. The window has two tabs: 'Concordance' (selected) and 'Statistics'. The interface is organized into several sections:

- Modify text:** A text input field for 'Resulting .txt file:' is followed by two buttons: 'Set File' and 'GO'.
- Extract units:** A 'Set File:' input field is followed by two buttons: 'Extract matching units' and 'Extract unmatching units'.
- Concordance presentation:** A checked checkbox labeled 'Use a web browser to view the concordance'. Below it is a button 'Show differences with previous concordance' and another button 'Show ambiguous outputs'.
- Show matching sequences in context:** This section contains three input fields: 'Context length: Left' (value 40), 'Context length: Right' (value 55), and 'Stop at:' (checkboxes for '{S}' and '{S}'). To the right is a dropdown menu 'Sort according to:' set to 'Center, Left'. A 'Build concordance' button is located at the bottom right of this section.

Concordance sur le motif <parler>

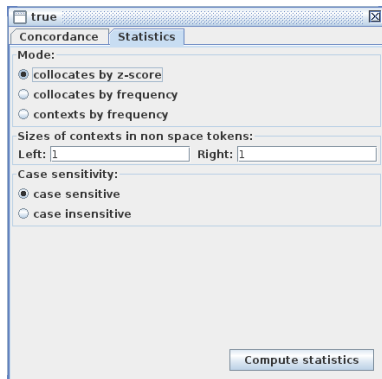


Concordance: /home/fortkare/unitex/French/Corpus/80jours_snt/concord.html

35 matches

x chevaux. {S} On partit. {S} Personne ne [parla](#) pendant le trajet, qui dura vingt minutes environ Fogg. {S} Pendant le jeu, les joueurs ne [parlaient](#) pas, mais entre les robes, la conversation à l'acception européenne du mot. {S} Elle [parlait](#) l'anglais avec une grande pureté, et le guide n Phileas Fogg, accoté dans son coin, ne [parlait](#) pas. {S} Passepartout, encore abasourdi, pressa " qui en disait long " ! mais il ne lui [parlait](#) pas, car il n'existait plus aucune intimité entre l'inspecteur de police, mais il ne lui [parlait](#) pas. {S} Depuis les derniers événements, leurs r apprit par le Figaro de l'endroit, qui [parlait](#) un assez bon anglais, que ces vieillards avaient ns communicatif que ce gentleman. {S} Il [parlait](#) aussi peu que possible, et semblait d'autant plus cile. {S} Il était évident que ce garçon [parlait](#) avec une absolue bonne foi, et qu'il n'était pas écouté Fix, et il fut convaincu que Fix [parlait](#) avec une entière bonne foi. " Sommes-nous amis à reconnaître ? {S} Et cependant, en lui [parlant](#) ainsi, Passepartout avait certainement eu une réflexion profonde et ne rêvait guère que l'on [parlât](#) de lui. " Le gouvernement anglais est extrêmement ONNAÎTRE DU TOUT LES CHOSES DONT ON LUI [PARLE](#) {S} Le Rangoon, l'un des paquebots qui rien, Arthémidore, Edrisi, ont toujours [parlé](#) avec épouvante, et sur lequel les navigateurs ne porte quel prix. " {S} Le capitaine avait [parlé](#) d'un ton qui n'admettait pas de réplique. " Mais t, un homme s'était approché, lui avait [parlé](#) même, mais l'agent l'avait renvoyé, après réponse ansport. " {S} C'était l'homme qui avait [parlé](#) à l'inspecteur de police pendant la nuit, et dont apître VIII {S} DANS LEQUEL PASSEPARTOUT [PARLE](#) UN PEU PLUS PEUT-ÊTRE QU'IL NE CONVIENT Français, qui ne pourra se retenir de [parler](#). {S} A bientôt, monsieur le consul. " {S} Cela dit eur. {S} D'ailleurs, il lui répugnait de [parler](#) à cet homme, dont il acceptait les services. {S} tes qu'il a manifesté l'intention de me [parler](#) ce soir ? _ Oui, madame. {S} Il s'agit sans doute _ Oui, pour l'instant. _ Alors venez me [parler](#). _ Que je... _ Dans l'intérêt de votre maître. " 'il n'était pas très difficile de faire [parler](#) ce garçon, et il se décida à rompre l'incognito agnon, et il n'essaya point de le faire [parler](#). {S} Une ou deux fois seulement, il entrevit Mr. er Fix dans ses idées. {S} Il fit encore [parler](#) le Français et acquit la certitude que ce garçon uns contre les autres, ne pouvaient se [parler](#). {S} Le froid, accru par la vitesse, leur eût cou e, même impassibilité. {S} Il resta sans [parler](#) pendant cinq minutes. {S} Puis levant les yeux su ulez-vous, monsieur Fix ? _ J'ai à vous [parler](#) de choses sérieuses. _ De choses sérieuses ! s'é cheveux en désordre. {S} .. Il ne pouvait [parler](#) ! " Monsieur, balbutia-t-il, monsieur... pardon. a chambre de Mr. Fogg. {S} Il ne pouvait [parler](#). " Qu'y a-t-il ? demanda Mr. Fogg. _ Mon maître. ers Colt. {S} Passepartout avait entendu [parler](#) de Sioux et de Pawnees, qui arrêtent les trains

Statistiques de collocations



Collocations du motif <parler>

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

Sources

Introduction

Concordancier

Recherche de motifs par expressions régulières

- Un pas plus loin

- Recherche de motifs référant aux dictionnaires

- Utilisation des méta motifs

Recherche de motifs par filtres morphologiques

Pour finir

Opérateurs

- ▶ **concaténation** :
 - ▶ point : $\langle DET \rangle . \langle N \rangle$ (reconnaît un déterminant suivi par un nom)
 - ▶ espace : $le \langle A \rangle chat$ (reconnaît l'unité lexicale *le*, suivie d'un adjectif et de l'unité lexicale *chat*)
 - ▶ note sur les parenthèses : servent de délimiteurs
- ▶ **union** :
 - ▶ + : $chat+chien \langle v \rangle$ (reconnaît l'unité lexicale *chat* ou *chien*, suivie par un verbe)
 - ▶ note sur epsilon : $le (petit+\langle E \rangle) chat$ (reconnaît les séquences *le chat* et *le petit chat*)
- ▶ **étoile de Kleene** : * (permet de reconnaître zéro, une ou plusieurs occurrences d'une expression)
 - ▶ $il\ fait\ très^* \text{ froid}$: 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 $\langle N \rangle^*$ produit-il une erreur? Que faire pour l'éviter?

Codes grammaticaux usuels

code	Signification	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

- ▶ $\langle E \rangle$: mot vide, ou epsilon. Reconnaît la séquence vide
- ▶ $\langle TOKEN \rangle$: n'importe quelle unité lexicale sauf l'espace
- ▶ $\langle MOT \rangle$: n'importe quelle unité lexicale formée de lettres
- ▶ $\langle MIN \rangle$: [...] de lettres minuscules
- ▶ $\langle MAJ \rangle$: [...] de lettres majuscules
- ▶ $\langle PRE \rangle$: [...] de lettres et commençant par une majuscule
- ▶ $\langle DIC \rangle$: n'importe quel mot figurant dans les dictionnaires du texte
- ▶ $\langle SDIC \rangle$: [...] mot simple [...]
- ▶ $\langle CDIC \rangle$: [...] mot composé [...]
- ▶ $\langle NB \rangle$: 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 >)
- ▶ ~ : **exclut** des codes (<A~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 »

Sources

Introduction

Concordancier

Recherche de motifs par expressions régulières

Recherche de motifs par filtres morphologiques

Définition

Exemples

Filtrer des motifs

Pour finir

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
- ▶ `<< ^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[^e]? >>` : contient ss suivi par un caractère qui n'est pas une voyelle e

Filtres plus complexes

- ▶ $\langle\langle [ai]ble\$ \rangle\rangle$: finit par *able* ou *ible*
- ▶ $\langle\langle ^{([rst][aeiouy])\{2,\}\$} \rangle\rangle$: 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 :

- ▶ $\langle V :K \rangle \langle \langle i\$ \rangle \rangle$: participe passé finissant par *i*
- ▶ $\langle CDIC \rangle \langle \langle .* \rangle \rangle$: mot composé contenant deux espaces
- ▶ $\langle A :fs \rangle \langle \langle ^pro \rangle \rangle$: 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*

Sources

Introduction

Concordancier

Recherche de motifs par expressions régulières

Recherche de motifs par filtres morphologiques

Pour finir

CQFR : Ce Qu'il Faut Retenir

TD



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

Illustrer la loi de Zipf

Exercice noté, à rendre avant le prochain cours, par mail

Construire, à l'aide d'Unitex et sur un tableur par exemple :

- ▶ deux (beaux) graphiques illustrant la loi de Zipf
- ▶ sur des textes de votre choix
- ▶ dans deux langues différentes
 - ▶ un bonus sera octroyé à qui traitera des langues peu courantes
 - ▶ si vous avez dû effectuer des traitements particuliers, signalez-les
 - ▶ si vous avez des commentaires à faire, faites-les

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*