Gestionnaire de Sélection et MOTIFS

Repérage et sélection de motifs multi-annotés (répétés)

01/12/2016 21:36:26

Serge Fleury

(document de travail)

Ce document présente la mise à jour mise en œuvre à partir de la **version 12.144** du *Trameur* permettant définir des MOTIFs puis de les mettre au jour via le *Gestionnaire de Sélection*.

1. Définitions

Source : (Née & al., 2017)

> Le **motif** se présente comme une nouvelle unité textuelle récurrente (A B C, par exemple) composé d'unités A, B, C, se situant à différents niveaux (formes graphiques, lemmes, catégories grammaticales, patrons syntaxiques et, éventuellement, schèmes métriques ou prosodiques). La première particularité du motif, en comparaison avec le segment répété ou la cooccurrence, est sa multidimentionnalité : « *la notion de motif est conçue comme un moyen de conceptualiser la multidimensionnalité (ou le caractère multi-niveau) de certaines formes récurrentes qui sollicitent à la fois le lexique, les catégories grammaticales et la syntaxe, éventuellement la prosodie, la métrique*. » (Longrée et Mellet 2013 : 66)

(à compléter)

2. Mise en œuvre dans le Trameur

Pour illustrer la fonctionnalité mise en œuvre dans le *Trameur*, on travaille ci-dessous avec le corpus suivant :

> Le corpus **VOEUX** réunit l'ensemble des allocutions du 31 décembre (vœux aux Français) des présidents de la Cinquième République, de 1959 à 2015. Il contient 58 257 occurrences pour 6 426 formes. Ce corpus est accessible en ligne sur les pages suivantes : (https://sourceforge.net/projects/txm/files/corpora/voeux/ et www.textopol.u-pec.fr [sur demande])

Ce corpus est chargé dans le *Trameur* avec étiquetage via *TreeTagger*.

Sur ce corpus, nous allons mettre au jour les séquences textuelles correspondant au motif suivant :

```
<2:nous> <1:avons;3:VER> <3:.*>{1,20} <2:espérer|espoir|confiance|raison|raisons>
```

Chaque item du motif est décrit dans une séquence du type <*description-item*>

La *description d'un item* contient des informations à vérifier sur une ou plusieurs couches d'annotation pour l'item visé; par exemple sur le second terme du motif on a : <1:avons;3:VER>, ce terme doit permettre de rechercher un VERbe dont la forme graphique est « avons » (l'ordre des annotations dans la *description d'un item* n'a aucune importance).

Le motif précédent décrit donc une séquence textuelle :

- commençant par un item dont l'annotation n°**2** (lemme) est *nous* (en gros le pronom personnel *nous* (minuscule ou majuscule si les lemmes sont tous codés en minuscule)
- suivi d'un item dont l'annotation n°**1** (forme) est *avons* et dont l'annotation n°**2** est VER ; plus précisément, un item dont l'annotation n°**1** contient la chaîne de caractère *avons* (par exemple, *avons* ou *savons*…) et dont l'annotation n°**2** contient la chaîne de caractère *VER*
- suivie de 1 à 20 occurrences d'items de valeur quelconque (via la regexp `.*`) pour l'annotation n°**3**
- et se terminant par des items dont l'annotation n°**2** (lemme) coïncident avec la *regexp* donnée *i.e* les unités suivantes : *espérer* ou *espoir*, ou *confiance* ou *raison(s)*

Ecriture des motifs dans un fichier

Tous les motifs doivent être écrits au préalable dans un fichier au format texte brut.

La figure suivante donne un voir un tel fichier ouvert via *Notepad++* :

modele-modifs.txt

```
PATTERN OK
Seul le pole doit etre decrit de maniere "exacte" dans sa forme
<1:patriotes>   #   <1:patriotes> <3:ADJ>
<1:patriotes>   #   <1:patriotes> <3:.*>{1,2} <3:ADJ>
<2:patriote>    #   <2:patriote> <3:PRP>{1,2} <3:ADJ>
<1:patriotes>   #   <1:patriotes> <2:.*>{1,5} <3:ADJ>
<1:patriotes>   #   <1:patriotes> <1:ont> <3:VER>{1,2} <3:NOM>
<2:patriote>    #   <2:patriote> <3:VER_pres> <3:VER>{1,2} <3:NOM>
<3:ADJ> #   <3:ADJ> <3:DET> <3:NOM> <3:VER>
<3:ADJ> #   <3:ADJ <3:.*>{1,2} <3:NOM>{2,3} <3:VER>
<2:le>  #   <2:le <3:.*>{1,2} <3:ADJ>{2,3} <3:NOM>
<3:DET_ART> #   <3:DET_ART> <3:ADJ>{0,2} <3:NOM> <3:VER>
<3:DET_ART> #   <3:DET_ART> <3:ADJ>{1,2} <3:NOM> <3:VER>
<3:DET_ART> #   <3:DET_ART> <3:ADJ> <3:NOM>{0,2} <3:VER>
<3:DET_ART> #   <3:DET_ART> <3:ADJ> <3:NOM>{1,2} <3:VER>
<3:DET_ART> #   <3:DET_ART> <3:ADJ>{1,2} <3:NOM>{1,2} <3:VER>
<3:DET_ART> #   <3:DET_ART> <3:ADJ>{1,2} <3:NOM>{1,2} <3:ADJ>{1,2} <3:VER>
<3:DET_ART> #   <3:DET_ART> <3:ADJ>{0,2} <3:NOM>{1,2} <3:ADJ>{1,2} <3:VER>
<3:DET_ART> #   <3:DET_ART> <3:ADJ>{1,2} <3:NOM>{1,2} <3:ADJ>{0,2} <3:VER>
<1:nous>    #   <1:nous> <2:avoir> <3:.*>{1,20} <2:espérer|espoir|confiance|raison|raisons>
<1:Nous>    #   <1:Nous> <1:avons> <3:.*>{1,20} <2:espérer|espoir|confiance|raison|raisons>
<2:nous>    #   <2:nous> <1:avons> <3:.*>{1,20} <2:espérer|espoir|confiance|raison|raisons>
<2:nous>    #   <2:nous> <1:avons;3:VER> <3:.*>{1,20} <2:espérer|espoir|confiance|raison|raisons>
<42:B>  #   <2:B> <42:I>{0,100} <42:L>
<1:patriotes;3:NOM> #   <1:patriotes;3:NOM> <3:PRP>{1,2} <3:ADJ>
<1:patriotes;3:NOM> #   <1:patriotes;3:NOM> <3:PRP>{1,2} <2:rencontrer;3:ADJ>
```

Un fichier similaire est fourni dans la distribution standard du logiciel.

Ce fichier a l'allure suivante :

POLE # MOTIF

Le POLE est suivi d'une tabulation, puis le caractère #, puis de nouveau une tabulation et le MOTIF. On trouve par exemple dans le fichier précédent la ligne suivante :

```
<3:DET_ART>     #     <3:DET_ART> <3:ADJ>{0,2} <3:NOM> <3:VER>
```

Elle contient la description du pôle (avant le #) puis celle du motif (après le #).

Le POLE correspond au premier terme du MOTIF :

3:DET_ART

ici un item dans l'annotation n°3 (catégorie) est DET_ART.

On a ensuite la description du MOTIF associé au pôle précédent :

<**3**:DET_ART> <**3**:ADJ>{0,2} <**3**:NOM> <**3**:VER>

Chaque item du motif est décrit dans une séquence du type <*description-item*> ; la *description d'un item* contient des informations à vérifier sur une ou plusieurs couches d'annotation pour l'item visé ; dans l'exemple précédent, seule une couche d'annotation est spécifiée pour chaque item du motif. Le motif testé *infra* en utilisera plusieurs.

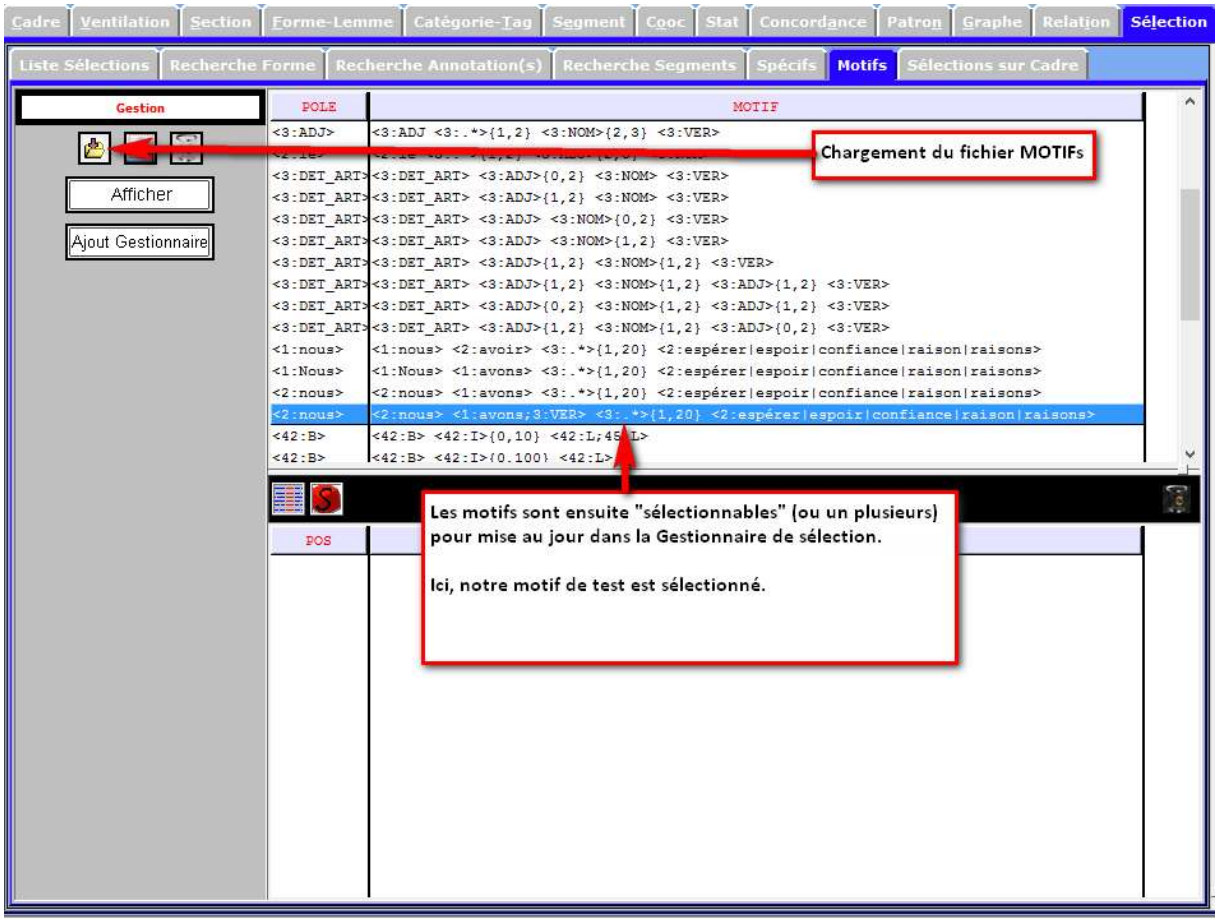
Le motif commence ici par le pôle précédent (<3:DET_ART>), il est suivi de 0 à 2 occurrences d'item(s) dont l'annotation n°3 (catégorie) est ADJ (<3:ADJ>{0,2}), puis d'un item dont l'annotation n°3 (catégorie) est NOM (<3:NOM>) et il se termine par un item dont l'annotation n°3 (catégorie) est VER (<3:VER>).

Actuellement, dans la description du motif, seul le premier terme du motif doit être défini précisément (dans sa valeur). Pour les autres termes du motif, il est possible d'utiliser une expression régulière (*cf* exemple initial).

Les termes du motif (sauf le premier et le dernier) peuvent être accompagnés d'un marquage de répétition (avec une valeur mini (0 compris) et une valeur maxi).

Chargement du fichier contenant les motifs

Une fois les motifs définis dans le fichier dédié, ce dernier peut être chargé dans le *Trameur* :



Remarque : le fichier peut être modifié puis rechargé pour éventuellement prendre en compte de nouveaux motifs.

Le motif sélectionné est celui explicité *supra* :

```
<2:nous> <1:avons;3:VER> <3:.*>{1,20} <2:espérer|espoir|confiance|raison|raisons>
```

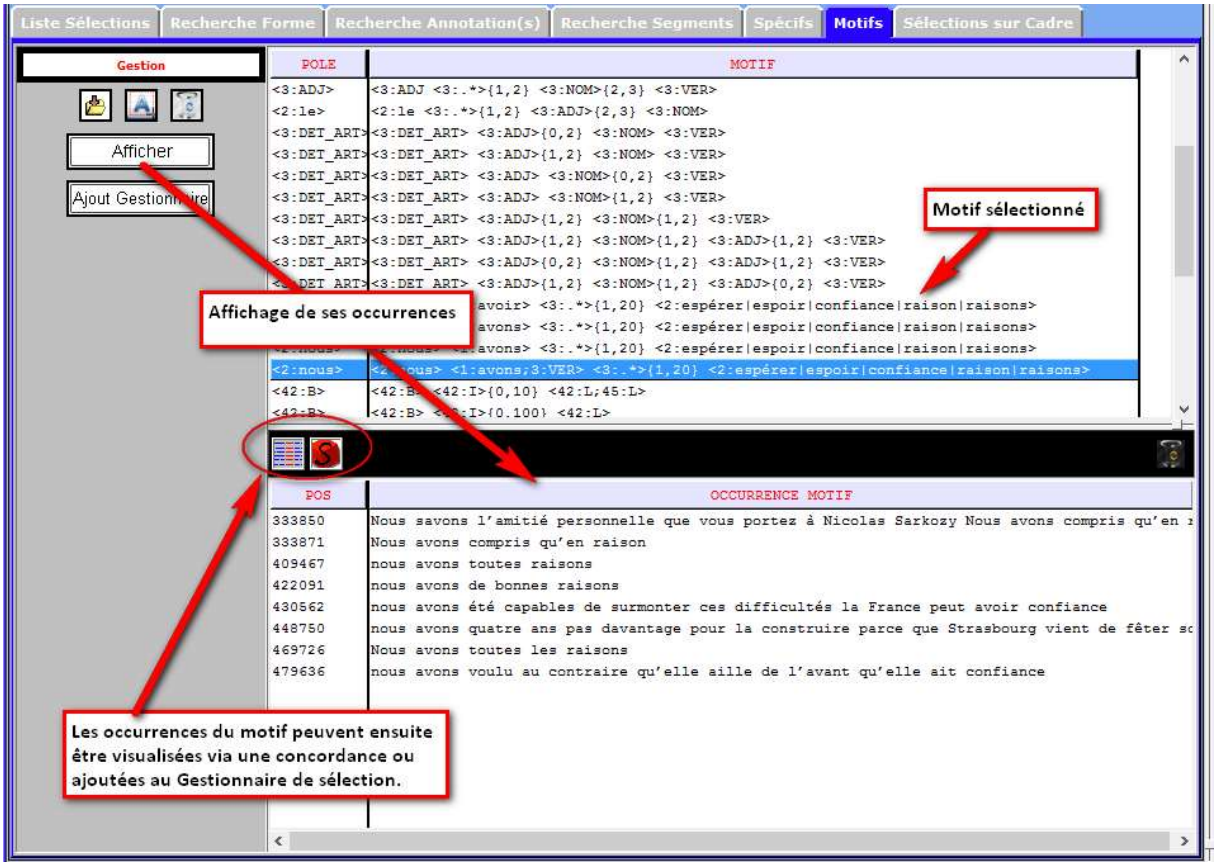
Rappel : dans ce motif, le second terme intègre une requête sur 2 couches d'annotations pour les items visés :
→ on cherche (1) le lemme « nous » suivi par (2) un item qui doit être un verbe ayant pour forme graphique « avons » etc.

Les différentes contraintes posées sur un item peuvent s'enchaîner les unes derrière les autres et séparées par un point-virgule et dans un ordre quelconque :

<i:*pattern_sur_i*;j:*pattern_sur_j*;k:*pattern_sur_k*;…>

1. Affichage des occurrences d'un motif

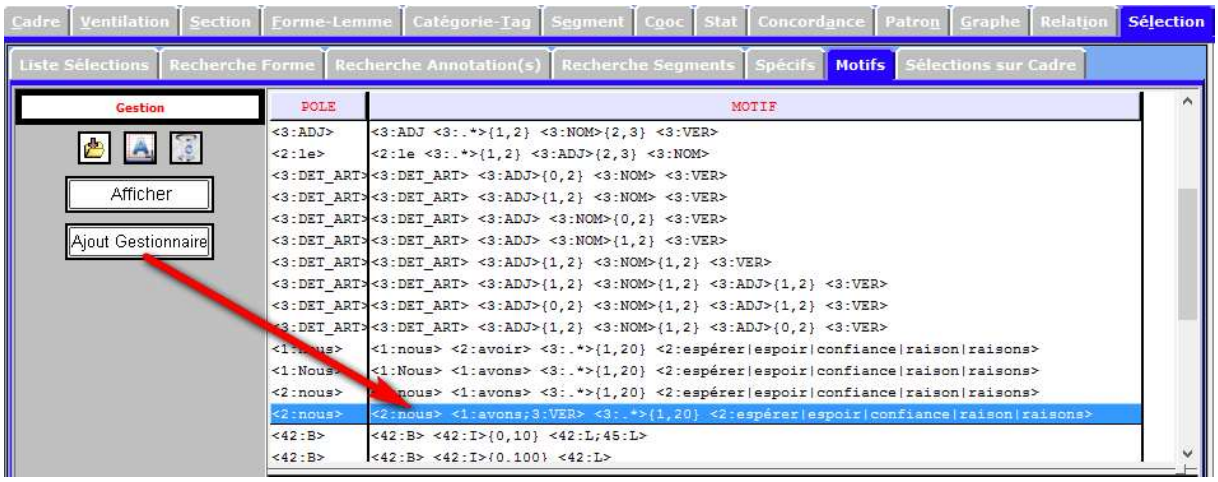
Le bouton « Affichage » permet d'afficher la séquence textuelle (les formes graphiques) des occurrences du motif :



Ces occurrences peuvent être affichées en contexte via une concordance : utile pour explorer les différentes annotations de chaque item de la séquence mise au jour.
Elles peuvent aussi être ajoutées au Gestionnaire de Sélection (sélection de la séquence visée puis activation du bouton « Sélection »).

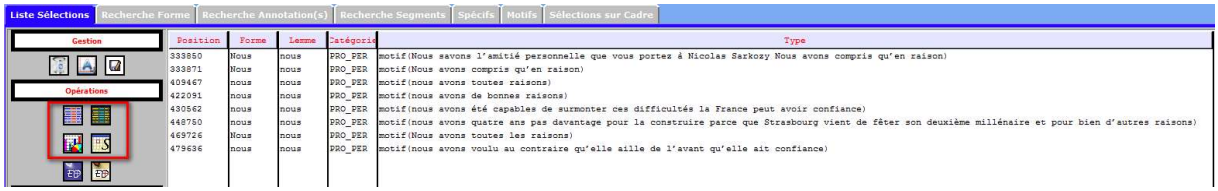
2. Sélection d'un motif à "mettre au jour" (*in fine* les occurrences du motif sont ajoutées au gestionnaire de sélection)

Le bouton « **Ajout Gestionnaire** » déclenche directement la recherche des occurrences des motifs sélectionnés et leur insertion dans le *Gestionnaire de Sélection* (sans passer par leur affichage comme précédemment)

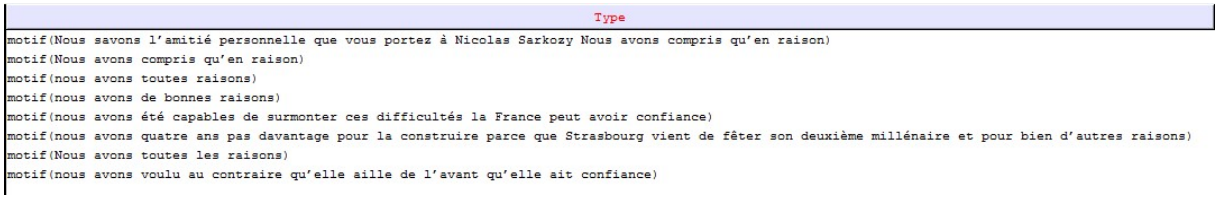


Affichage du résultat dans le Gestionnaire de Sélection

Une fois les occurrences du motif visé éventuellement ajoutées au Gestionnaire de Sélection, il est possible de leur appliquer les traitements disponibles.



La quatrième colonne donne à voir les séquences textuelles associées au motif (via leur annotation n°1 (forme)) :

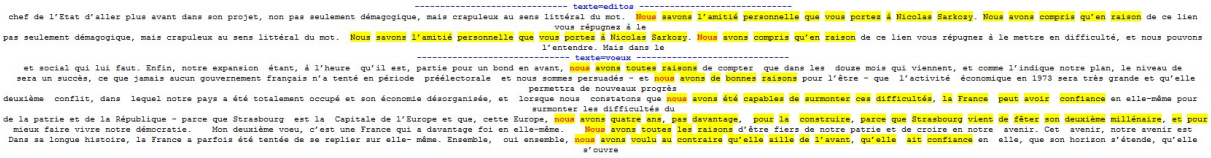


REMARQUE : la recherche de motif "traverse" les délimiteurs....

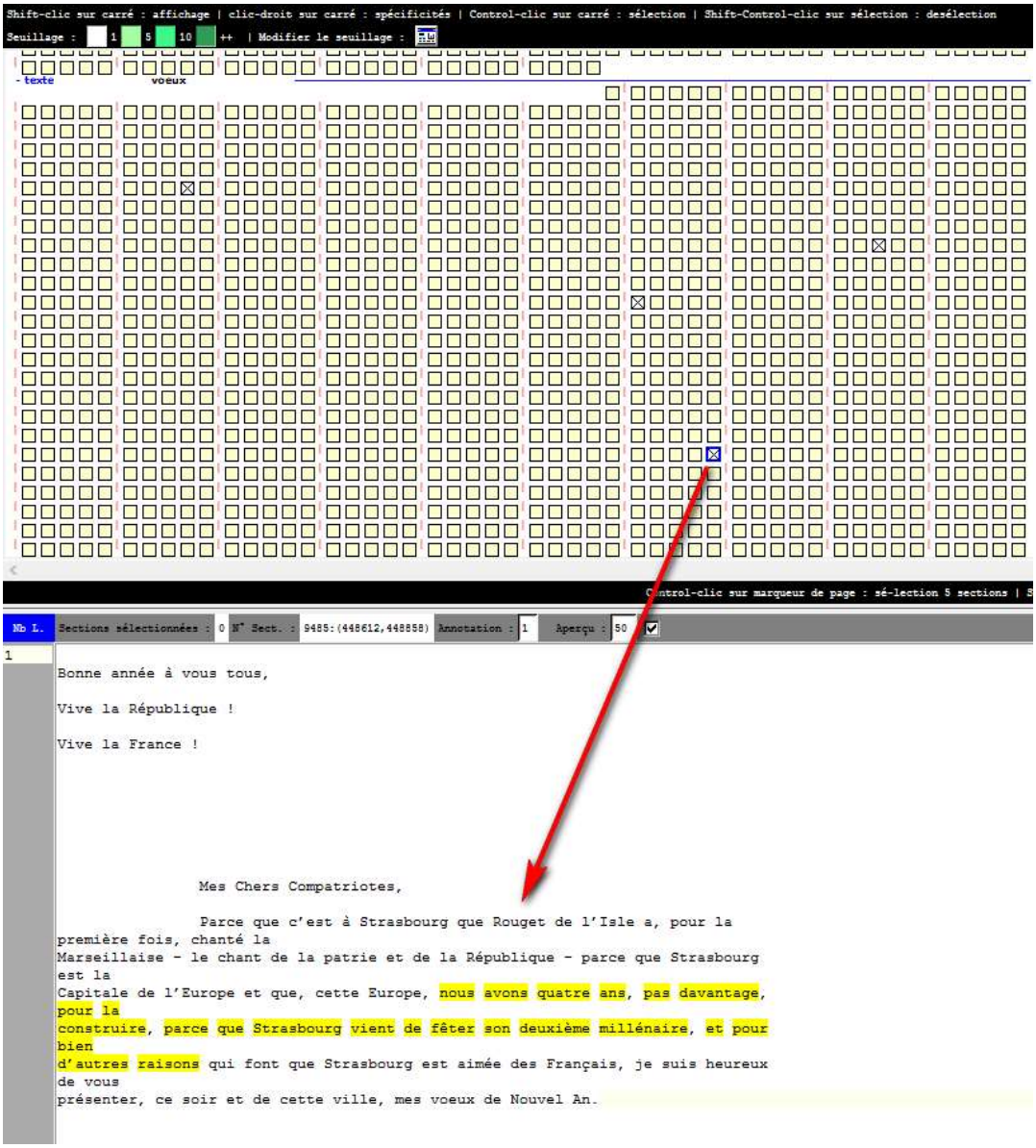
Visualisation en contexte du "motif"

Les items ainsi mis au jour (et les séquences associées : les occurrences du MOTIF) sont ensuite exploitables via les opérations disponibles dans le logiciel :

On peut visualiser ces séquences en contexte, via une concordance :



ou dans une section de la *Carte des Sections* :



Ou encore dans un graphique de ventilation (projection ici de toutes les occurrences du motif) :