Notes préparatoires

pour la réunion du 18 juin 2018

sur mon projet de recherche

**Sommaire**  
**1. Les différentes possibilités après un « : »**   
**2. Description et assimilation des différentes possibilités à l'aide de patrons**  
**3. Limites à l'assimilation des possibilités**  
**4. Les exemples d'une combinaison  
5. Pistes futures  
  
1. Les différentes possibilités après un « : »**   
  
À partir de mon corpus lemmatisé et étiqueté en parties du discours (POS), j'ai voulu savoir ce que l'on pouvait trouver comme combinaisons de POS sur les 5 premiers éléments suivant un « : » dans un titre. Le premier fichier (3a) liste les 7 319 possibilités.  
  
Ce premier résultat m'a permis de déterminer la combinaison la plus répandue (6% du total des titres avec un « : ») soit : DET   NC   P    DET  NC     
Soit un déterminant suivi d'un nom commun, suivi d'une préposition, suivi d'un déterminant, suivi d'un nom commun.   
  
Les 6 premières combinaisons :  
1.    DET   NC   P    DET  NC    5346    6%  
2.    NC    P    DET  NC   P     2422    3%  
3.    NC    P    DET  NC   ADJ   1816    2%  
4.    DET   NC   P+D  NC   P     1491    2%  
5.    NC    CC   NC   -    -     1450    2%  
6.    DET   NC   ADJ  P    DET   1438    2%  
  
Les codes des POS sont fournis dans le tableur et repris directement de Talismane.  
-     signifie absence d'éléments (pas de forme, lemme et POS).  
  
**2. Description et assimilation des différentes possibilités à l'aide de patrons**  
  
Pour regrouper certaines possibilités entre elles, nous nous étions posés la question d'une approche automatique ou humaine.  
L'approche humaine est pour l'instant la plus simple :  
  
1) Tout d'abord, on lève la limite arbitraire de 5 éléments pour dire que l'on cherche juste à associer ce qui suit un « : » dans un titre avec un patron.  
  
2) On écrit une première fois le patron P1 ainsi : DET?  NC   P    DET  NC      
    où « ? » marque le caractère optionnel du premier déterminant  
  
On voit bien que les lignes 1, 2 et 3 correspondent à ce patron.  
  
3) On complète notre patron P1 ainsi : DET?  NC   (P    DET  |P+D  )NC      
    où | propose un choix entre P    DET   et P+D  .  
  
On voit bien que les lignes 1, 2, 3, 4 correspondent à ce patron.  
La ligne 4 est assimilée en offrant le choix entre préposition simple et une préposition combinée avec un déterminant  
  
4) On complète notre patron P1 ainsi : DET?  NC   ADJ  ?(P    DET  |P+D  )NC      
La ligne 6 est assimilée en offrant la possibilité d'avoir un adjectif après le premier nom.  
  
Les lignes 1, 2, 3, 4 et 6 pouvant être assimilée à un même patron, P1.  
Seule la ligne 5, qui est NC    CC   NC   -    -    , est vraiment dissemblable.  
On peut dire qu'elle correspond au patron P2 : NC    CC   NC     
  
**3. Limites à l'assimilation des possibilités**  
  
On peut se poser la question jusqu'où doit-on complexifier le patron pour assimiler les lignes.  
En effet, il serait possible en reprenant mon mini-langage de construire un patron qui assimilerait toutes les lignes présentées :  
(DET?  NC   ADJ  ?(P    DET  |P+D  )NC  |NC    CC   NC   )  
  
Je ne pense pas que cette dernière complexification soit une bonne idée car nous n'avons pas affaire à la même structure syntaxique.  
À droite nous avons une coordination entre deux noms de même niveau, alors qu'à gauche les noms ne sont pas sur le même niveau.  
  
Par contre, on peut avoir des patrons imbriqués.  
Si on réécrit notre patron P1 en rendant optionnel le complément de nom : DET?  NC   ADJ  ?((P    DET  |P+D  )NC   )?  
On peut réécrire P2 pour le rendre plus flexible ainsi : P1    CC   P1       
  
Pour pouvoir détecter qu'un titre se conforme à P2, il est alors nécessaire d'aller voir au-delà des 5 premiers éléments après le « : » et au-delà du premier patron que l'on arrive à associer.  
Autrement dit, il faut essayer d'associer le plus d'éléments possibles au patron le plus complexe possible.  
  
**4. Les exemples d'une combinaison**  
  
On peut ensuite s'interroger sur les lemmes utilisés sous ces POS.  
  
Pour la combinaison DET   NC   P    DET  NC   , j'ai voulu voir l'ensemble des 5 346 quintuplés de 5 lemmes/formes qui correspondent à cette possibilité.  
  
C'est le second fichier (3b).  
  
En faisant un calcul de fréquences sur le premier nom (NC1), on a les lemmes suivants avec une fréquence > 100 :  
cas           701  
exemple       356  
apport        128  
approche      118  
enjeu         110  
  
En faisant un calcul de fréquences sur le second nom (NC2), on a les lemmes suivants avec une fréquence > 50 :  
\_             242  
littérature    60  
histoire       54  
épreuve        53  
recherche      51  
  
(Il faut que j'enquête sur le "\_" )  
  
**5. Pistes futures**  
Voici quelques pistes :  
  
- Décrire entièrement mon langage de description de patrons  
- Établir une liste des patrons  
- Construire l'algorithme d'association d'un titre avec un ou des patrons  
- Refaire un calcul de fréquence mais avec les différents patrons cette fois-ci  
- Faire des fréquences de NC1 pour P1 (= "lemme préféré")  
- Faire la fréquence des couples (NC1, NC2) pour P1  
- Faire la fréquence des couples (NC1, NC2) pour P2  
- Trouver les lemmes préférés de chaque patron