Zeina Tmart (ENS de Lyon)
Alexei Lavrentiev (CNRS - IHRIM)
Céline Guillot-Barbance (ENS de Lyon)
Sophie Prévost (CNRS - Lattice)

Atelier Diachro X

jeudi 2/06/2022

Outils pour l'exploration lexicale, morphosyntaxique et syntaxique de la Base de français médiéval

Programme :

14h-15h30

1. Lemmes et étiquettes morphosyntaxiques (requêtes CQL)
   Base de français médiéval : https://txm.bfm-corpus.org
2. Requêtes syntaxiques (CQP)
   Corpus SRCMF2022 en ligne : https://txm-bfm.huma-num.fr

Pause

16h-17h30

3. Requêtes syntaxiques (TIGERsearch)
   Corpus SRCMF2022 avec le logiciel TXM pour poste
4. Annotation de corpus
   Corpus ATELIER-DIACHRO avec le logiciel TXM pour poste

1. Lemmes et étiquettes morphosyntaxiques (requêtes CQL)

Portail de la Base de français médiéval : txm.bfm-corpus.org

Corpus : BFM2019 et BFM2019 lemmatisé

Jeu d'étiquettes Cattex 2009 : http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_2.0.pdf

- **VER**
  - cjg
  - **inf**
  - ppe
  - ppa
- **PRO**
  - per
  - imp
  - adv
  - pos
- **DET**
  - def
  - ndf
  - ind
  - pos

- **dem**	- **dem**
- ind	- car
- rel	- rel
- int	- int
- com	- com

1.1. Formuler une requête à partir des propriétés de mots

- Commande **Index** :
  - Afficher tous les verbes à l'infinitif
    - Requête : `[cattex-pos="VERinf"]`
  - Cibler la requête sur le verbe *faire*
    - Requête : `[cattex-pos="VERinf" & word="f..re"]`
  - Afficher les propriétés *lemma* et *lemma_src*
  - Éliminer le bruit
    - Requête : `[cattex-pos="VERinf" & word="ff?[ae]i?re"]`

1.2. Recherche sur une succession de deux propriétés

- Commande **Index** :
  - Chercher les déterminants et pronoms démonstratifs
    - Requête : `[cattex-pos= ".*dem"]`
  - Chercher quels mots suivent les déterminants démonstratifs
    - Requête : `[cattex-pos="DETdem"][]`
  - Préciser la requête pour viser les pronoms relatifs subséquents
    - Requête : `[cattex-pos="PROdem"] [cattex-pos="PROrel"]`

1.3. La propriété "q" : rechercher les démonstratifs en discours direct

- Commande **Lexique** :
  - Afficher la propriété "q":
    - 0 : hors DD
    - 1 : DD de niveau 1
    - 2 : DD de niveau 2
    - 3 : DD de niveau 3
- Commande **Index** :
  - Afficher la propriété "q"
    - Requête : `[lemma contains "(cil|cist)"]`
- Commande **Concordance** :
  - Les démonstratifs en DD de niveau 3
    - Requête : `[q="3" & lemma contains "(cist|cil)"]`

1.4. Exercice :

- Commande **Concordance** :

- Chercher un lemme “cil” , “cist” ou “ce” dans un DD de niveau 1
  - Trier les résultats sur le pivot
  - Trier les résultats sur le contexte gauche
- Chercher les incises de type “ce dist X”
- Chercher les incises de type “ce fait X”

2. Requêtes syntaxiques (CQL)

Portail txm-bfm (expérimental) : https://txm-bfm.huma-num.fr

- Se connecter
  - nom d’utilisateur : diachro10
  - mot de passe : diachro10

Corpus : SRCMF2022 (11 textes, 200 519 mots et ponctuations)

Requêtes à copier-coller : https://bit.ly/3N1tFoX

- Enregistrer le fichier requetes-syntaxiques.txt dans le dossier de travail

Liste des propriétés de mots :

- ud-deprel = relation de dépendance (https://universaldependencies.org/u/dep/index.html)
  - [root = racine de la phrase (verbe conjugué, participe ou nom)]
  - aux = verbe auxiliaire
  - nsubj = sujet nominal
  - obj = objet direct
  - iobj = objet indirect
  - obl = oblique
  - nmod = modifieur nominal
  - det = déterminant
  - conj et cc = conjonctions de subordination et de coordination
  - ccomp et xcomp
- ud-head-deprel = relation de dépendance du mot gouverneur
- ud-dep-deprel = relations de dépendance des mots gouvernés
- ud-id = identifiant (position du mot dans la phrase)
  - [ud-id!="1"]* pour rester dans la même phrase

2.1. Principes de l’annotation syntaxique UD

- Afficher l’étiquette syntaxique d’un pronom démonstratif
  - Commande **Index**
    - Afficher les propriétés *lemma* et *ud-deprel*
    - Requête `[cattex-pos="PROdem"]`
- Pour le déterminant démonstratif, on s’intéresse à sa « tête »
  - Commande Index
  - Afficher les propriétés *lemma* et *ud-head-deprel*
  - Requête `[cattex-pos="DETdem"]`

2.2. Recherche de phrases OSV (proposition principale)

Recherche de phrases OSV (proposition principale)

a. Verbe conjugué

- Construction de la requête
  - Objet `[ud-deprel="obj"]` →
    - + dépend du verbe principal
      `[ud-deprel="obj" & ud-head-deprel="root"]`
  - 0 ou plusieurs mots `[]*` →
    - + sans traverser la frontière de la phrase
      `[ud-id!="1"]*`
  - Sujet `[ud-deprel="nsubj" & ud-head="root"]` →
    - + dans la même phrase !
      `[ud-id!="1" & ud-deprel="nsubj" & ud-head="root"]`
  - 0 ou plusieurs mots dans la même phrase `[ud-id!="1"]*`
  - Verbe de la même phrase
    `[ud-id!="1" & ud-deprel="root" & cattex-pos="VERcjg"]`
    - en UD le verbe auxiliaire dépend de la racine nominale ou participiale

```
[ud-deprel="obj" & ud-head-deprel="root"][ud-id!="1"]* [ud-id!="1"
& ud-deprel="nsubj" & ud-head-deprel="root"] [ud-id!="1"]*
[ud-id!="1" & ud-deprel="root" & cattex-pos="VERcjg"]
```

b. Verbe auxiliaire

- Pour retrouver les phrases avec un verbe auxiliaire, on modifie le dernier élément
  - Auxiliaire qui dépend de la racine de la phrase :

```
[ud-deprel="obj" &
ud-head-deprel="root"][ud-id!="1"]*[ud-deprel="nsubj" &
ud-head-deprel="root" & ud-id!="1"] [ud-id!="1"]*
[ud-deprel="aux" & ud-id!="1" & ud-head-deprel="root"]
```

c. Les deux

- Pour retrouver les deux, on combine les conditions avec l'opérateur « | » (OU) et des parenthèses

```
[ud-deprel="obj" & ud-head-deprel="root"] [ud-id!="1"]*
[ud-deprel="nsubj" & ud-head-deprel="root" & ud-id!="1"]
[ud-id!="1"]*
[ud-id!="1" & (( ud-deprel="root" & cattex-pos="VERcjg" ) |
(ud-deprel="aux"  & ud-head-deprel="root"))]
```

3. Requêtes syntaxiques (TIGERsearch)

3.1. Préparation

- Configurer TXM

    - passer au niveau de mise à jour ALPHA
        - Editer > Préférences > TXM > Avancé > Niveau de mise à jour > sélectionner ALPHA
            - Cliquer sur « Apply and close »
    - installer l'extension "TIGERSearch"
        - Fichier > Ajouter une extension > TIGERSearch
            - Accepter les choix proposés
    - mettre à jour TXM et l'extension
        - Fichier > Vérifier les mises à jour
        - Accepter les options proposées pour TXM et TIGERSearch
    - Ajouter le moteur TIGER aux concordances
        - Editer > Préférences > TXM > Avancé > Search engines
            - Cocher « Show available search engines »
- Télécharger le corpus SRCMF2022 : https://bit.ly/3MBJCIt
- Charger le corpus dans TXM
    - Fichier > Charger > SRCMF2022-2022-05-20.txm

3.2. Langage de requêtes TIGERSearch

- Langage similaire à CQL
- Requêtes sur plusieurs lignes
    - définition de variables : #obj:[cat="obj"]
    - opérateur & : combiner des conditions
    - // : pour ajouter un commentaire
    - >D : relation de dépendance syntaxique
    - >L : expression lexicale
- Nœuds terminaux et non terminaux
- Le moteur retourne un ensemble de phrases qui correspondent à la requête
    - plusieurs "matchs" possibles dans une phrase

3.3. Exemple de requête : phrases OSV

```
#pivot:[pos="VERB"]
& #clause:[cat="root" & type="VFin"]
& #clause >L #pivot
& #clause >D
#obj:[cat=("obj"|"ccomp"|"obj\:advneg"|"obj\:advmod")]
& #clause >D #suj:[cat=("nsubj"|"csubj")]
& #obj     >L #objhead:[]
& #suj     >L #sujhead:[]
& #objhead .* #sujhead & #sujhead .* #pivot //OSV//
```

- Retrouvez cette requête (Requête 4))dans le fichier requetes-syntaxiques.txt et copiez-la

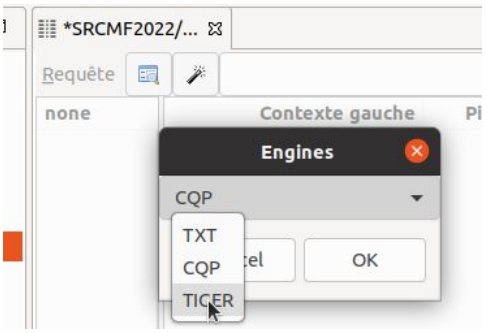
3.4. « Arbres syntaxiques » (portail TXM-BFM et TXM 0.8.1)

- Portail https://txm-bfm.huma-num.fr/txm/
    - cliquer sur le corpus SRCMF2022, puis sur l'icone « tête de tigre »

    - coller la requête, puis cliquer sur « Chercher »
- TXM 0.8.1
    - Commande « Arbres syntaxiques »
    - même manipulation que sur le portail

3.5. Concordance TIGER (TXM 0.8.1)

- Sélectionner le corpus SRCMF2022
- Utiliser la commande « Concordances »
    - sélectionner le moteur TIGER
    - Copier-coller la Requête 5 depuis le fichier «requetes-syntaxiques.txt »




```
#pivot:[pos="VERB"] & #clause:[cat="root" & type="VFin"] & #clause
>L #pivot & #clause >D
#obj:[cat=("obj"|"ccomp"|"obj\:advneg"|"obj\:advmod")] & #clause
>D #suj:[cat=("nsubj"|"csubj")] & #obj     >L #objhead:[] & #suj >L
#sujhead:[] & #objhead .* #sujhead & #sujhead .* #pivot //OSV//
```

3.6. Exercice:

- Modifier la requête pour retrouver
    - les phrases SOV
    - les phrases attributives, avec un sujet postposé au verbe auxiliaire

4. Annotation de corpus

Corpus ATELIER-DIACHRO : https://bit.ly/3NxjA2C

4.1. Corriger des lemmes par l'annotation

- Commande **Index** :
    - Requête : `[lemma contains "f[ae]i?re"%cd]`
- Commande **Concordance** :
    - Double-clic
    - Requête : `[word="ferè" & lemma="\|ferè\|"]`

- Procédure d'annotation :
    - Insérer un "@" dans la requête pour marquer le pivot
    - Sélectionner le crayon d'annotation
    - Choisir la propriété *lemma*
    - Corriger le lemme => |faire|
    - Sauvegarder l'annotation

4.2. Exercice : créer de nouvelles propriétés

- Temps => tpsV
    - “fist”
    - “face”
- Personne => persV
    - “fait”