

### Objectifs

- Prendre en main le logiciel TXM
- Utiliser de nouveaux corpus avec TXM
- Etiqueter morpho-syntaxiquement
- Faire des requêtes complexes

NB: Ce TD sera évalué pour votre note de contrôle continu, vous devrez répondre à des questions sur Moodle.

### Exercice 1 : Requêtage, tri et analyse de Concordances

#### Requête et tri

Toujours sur le corpus Vœux, observer les concordances du mot "temps".

- Triez selon le contexte droit pour faire apparaître les co-occurents droits.
- Triez selon le contexte gauche : comment sont filtrés les résultats ?
- Faites apparaître l'année et le locuteur correspondant à chacune des occurrences : clic droit sur text.id → options d'affichage des références

#### Contextes des occurrences

- Accédez au texte complet d'une occurrence de "temps" (clic droit puis afficher en plein texte) afin d'observer son contexte complet, on parle aussi de **retour au texte**.
- Observez que vous pouvez manipuler les différentes fenêtres. Mettez la fenêtre plein texte en regard de la liste des concordances pour faciliter les comparaisons.
- Modifiez la fenêtre d'affichage des concordances (taille du contexte gauche et du contexte droit).
- Affichez les concordances du nom "jeune" dans le corpus (singulier et pluriel) grâce aux requêtes CQL
- Modifiez l'affichage du pivot pour faire apparaître l'étiquette POS dans les résultats (cf. Figure 1)
- Affichez les formes verbales de "pouvoir" sans les formes nominales

### Exercice 2 : Fonctionnalités supplémentaires

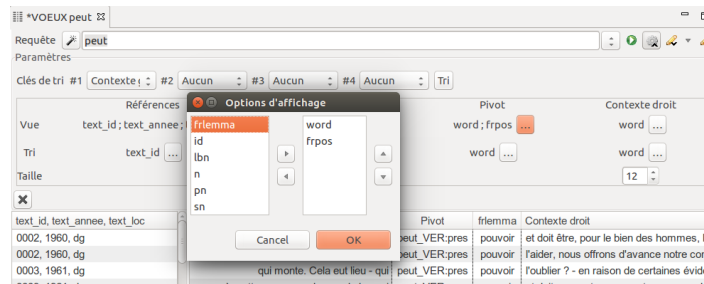


Figure 1: Ajout de propriétés dans l'affichage du pivot

## Comprendre une structure Xml

Récupérez l'échantillon d'articles de l'Est Républicain Décompressez l'archive dans un dossier corpus. Chaque fichier XML correspond à une édition du journal.

Ouvrez dans un éditeur de texte le fichier `Corpus/Annee1999/1999-05-17.xml`, observez la structure.

## Import Xml simple dans Txm

- Menu Fichier puis importer, choisissez le format XML/w +csv.
- Sélectionnez le dossier `Corpus/Annee1999` qui contient un échantillon d'articles en Xml
- Dans la partie "Langue principale" (cf. Figure 2), sélectionnez "fr" (si vous ne le faites pas, rien n'apparaîtra dans le menu contextuel après l'import).
- Lancez l'import du Corpus



Figure 2: Le menu import

ANNEE1999 devrait apparaître dans le menu de gauche

Observez les propriétés de ce Corpus ANNEE1999. Sont-elles renseignées ? Pourquoi d'après vous ?

Le corpus n'est pas enrichi, TXM ne peut donc afficher que des choses assez simple. Observez dans le menu index via la baguette magique (Figure 3) que vous n'avez accès qu'à des informations limitées sur les mots (Figure 4).



Figure 3: L'icône de la baguette magique

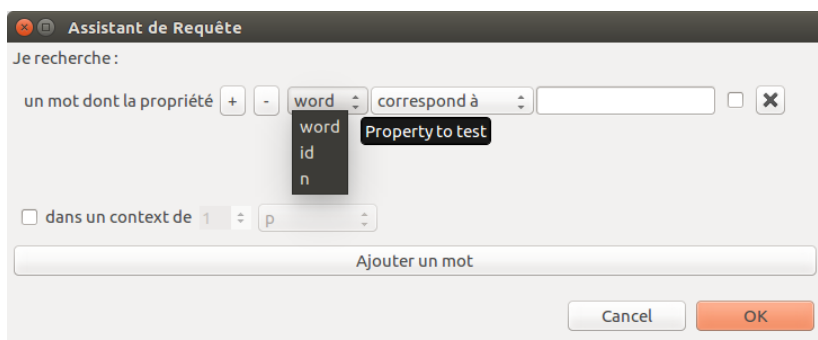


Figure 4: Propriétés des mots sans enrichissement

## Quelques recherches

Recherchez dans le corpus (via la fonction cooccurrences) les voisinages les plus fréquents des noms de villes suivants :

- Epinal
- Bar-le-Duc
- Toul

Cherchez une explication aux concordances fréquentes trouvées, pourquoi a-t-on souvent des parenthèses ... ?

Comparez pour chacune des villes vos hypothèses avec les concordances que vous observez via la fonction concordances.

Observez maintenant les cooccurrences de "Nancy", ne considérez pas les mots dont la fréquence est supérieure à 400 (de manière à faire disparaître les ponctuations et les mots outils).

Observez les cooccurrences dont l'"Indice" est supérieur ou égal à 15. Rangez les par catégories (rues, sigles ...). NB: utilisez la fonction "concordance" (accessible en cliquant droit sur le mot) et/ou Internet pour vérifiez que vous mettez bien les mots dans la bonne catégorie.

## Enrichir le Corpus

Avec le menu Fichier puis ajouter une extension, ajoutez le module Treetagger (deux éléments : base et modèles).

Vous allez ré-importer le corpus mais cette fois en activant l'option "Annoter le corpus" (cf. Figure 2).

Treetagger va avoir lemmatisé et étiqueté morpho-syntaxiquement le corpus, vous pouvez le voir en allant dans la fonction index (cf. Figure 5).

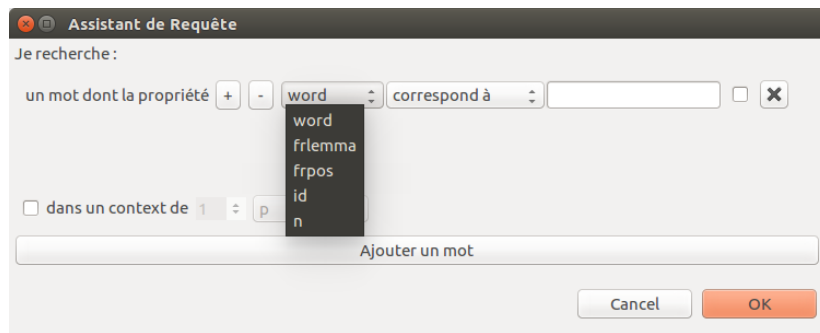


Figure 5: Propriétés des mots après enrichissement

La liste des étiquettes utilisées par Treetagger figure ici : <https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html>.

Utilisez l'index pour chercher les noms propres les plus fréquents (l'étiquette est NAM).