Exploration informatique de corpus annotés
Projet final :
Recherche et extraction de déclencheurs de présupposition en corpus

L'objectif du projet est de réaliser une « grammaire locale » (au sens où ce terme est employé dans l'application Unitex) pour repérer, dans des textes littéraires, le plus grand nombre possible d'occurrences de déclencheurs de présupposition.

> Les déclencheurs de présuppositions sont des lexèmes ou des constructions syntaxiques qui ont la capacité de produire (« déclencher ») des présuppositions. Par exemple les verbes factifs comme *regretter*, ou les constructions clivées (*c'est X qui P*). Rechercher les occurrences de ces expressions ou constructions qui ont un effet présuppositionnel se heurte à deux types de difficultés. La première est commune à toute étude de corpus : il faut s'assurer que la suite de caractères que l'on repère est bien une instance du lexème recherché, dans l'emploi recherché. Ainsi, si on cherche le lexème *voir* il faut garder à l'esprit que la forme fléchie *vis* peut aussi correspondre au lexème *vivre* ; si on cherche l'adverbe négatif *plus*, il faut garantir qu'on ne le confond pas avec le comparatif *plus*. De même, il faut vérifier que le contexte syntaxique correspond à un emploi pertinent du lexème. Par exemple le verbe *regretter* n'est présuppositionnel que quand il a une complétive. Le fait que le corpus soit annoté constitue une aide à condition de garder en tête que les annotations automatiques présentent toujours un taux d'erreur non négligeable.
>
> La seconde difficulté, spécifique au phénomène de la présupposition, est en lien avec le fait que certains contextes « suspendent » ou « annulent » les présuppositions. Par conséquent, il peut arriver qu'une occurrence bien identifiée d'un lexème présuppositionnel, dans un contexte spécifique, ne déclenche pas de présupposition. Par exemple, alors que l'adverbe aspectuel *encore* est considéré comme un déclencheur de présupposition, nous avons observé que lorsqu'il est combiné avec le comparatif *plus*, il ne semble jamais déclencher de présupposition, et même si ce fait reste à expliquer, il a conduit à ne pas annoter comme déclencheur les occurrences de *encore plus* en corpus.
>
> Dans le cadre du présent projet, le premier type de difficulté sera nécessairement abordé, et le rapport devra décrire les problèmes rencontrés et les choix mis en œuvre ; le traitement, même partiel, du second type de difficulté sera considéré comme un plus.

Chaque groupe de 2 ou 3 personnes choisit une classe de déclencheurs dans la liste suivante. Ces classes sont définies succinctement et illustrées dans l'article donné en référence. Plusieurs groupes peuvent choisir la même classe, même si l'idéal serait de bien les répartir.

- Prédicats Aspectuels
- Adverbes Restrictifs
- Subordonnées Causales
- Subordonnées Temporelles
- Prédicats Factifs
- Constructions Clivées
- Adverbes Additifs

La constitution du groupe doit être définitive au plus tard le 27 avril 2021 .

Méthode de travail

1. Dans un premier temps, on demande donc de réaliser un automate dans Unitex qui permette de retrouver les occurrences des déclencheurs de la classe choisie dans le texte de Jules Verne qui a servi pour l'article de référence ("De la Terre à la Lune"). Il est plus facile de réaliser d'abord un automate par déclencheur, avec l'objectif de repérer le même nombre (ou un nombre proche) d'occurrences que celles qui sont indiquées dans l'article (ce qui garantit jusqu'à un certain point la méthode). Ensuite ces automates peuvent être regroupés pour former une "grammaire locale", qui peut identifier et extraire toutes les occurrences de la classe.

2. Dans un second temps, on choisira un nouveau texte de Jules Verne, et on cherchera à appliquer l'automate trouvé pour repérer et extraire les occurrences de déclencheurs de présupposition du nouveau texte. On peut s'attendre à ce qu'il y ait du bruit (des choses trouvées que l'on ne souhaite pas extraire) et du silence (des choses qu'on aurait du extraire mais qui « échappent » à la grammaire locale). Il faut donc profiter de ce travail sur un nouveau texte pour raffiner la grammaire locale, l'objectif (idéal) étant qu'elle devienne suffisamment précise et générale pour pouvoir être appliquée avec confiance sur n'importe quel texte.
3. Les occurrences extraites seront mise en forme dans un tableur, idéalement sous la forme KWIC, et éventuellement filtrées manuellement.

Rendu

— On demande un rapport (4 à 6 pages hors annexes) qui décrit les étapes du projet, les difficultés rencontrées et les choix effectués.
— On demande aussi les fichiers présentant la ou les grammaires locales (graphes) élaborés.
— Enfin, on demande un fichier excel/csv contenant les données extraites sur le nouveau texte.
— Sur la base d'une "analyse d'erreurs" des données extraites sur le nouveau texte, on pourra proposer des pistes d'amélioration du processus, que ce soit au niveau de la grammaire locale ou de l'annotation du corpus.
— L'ensemble des éléments doivent être déposés sur iCampus au plus tard le 14 mai 2021.

27 avril 2021 (minuit)	Constitution du groupe	à indiquer dans le tableur pointé sur iCampus.
27 avril 2021 (minuit)	Choix du 2e corpus	*idem*
14 mai 2021 (minuit)	Rendu final du projet	à déposer sur iCampus Formats acceptés : pdf (rapport) ; .grf (graphes unitex) ; xls/csv (données extraites)