Description et modélisation des chaînes de référence. Le projet ANR Democrat (2016-2020) et ses avancées à mi-parcours

Frédéric Landragin, Marine Delaborde, Yoann Dupont, Loïc Grobol

Comme tout avait brûlé – la mère, les meubles et les photographies de la mère -, pour Fabre et le fils Paul c'était tout de suite beaucoup d'ouvrage : toute cette cendre et ce deuil, déménager, courir se refaire dans les grandes surfaces. Fabre trouva trop vite quelque chose de moins vaste, deux pièces aux fonctions permutables sous une cheminée de brique dont l'ombre donnait l'heure, et qui avaient ceci de bien d'être assez proches du quai de Valmy.

Le soir après le dîner, Fabre parlait à Paul de sa mère, sa mère à lui Paul, parfois dès le dîner. Comme on ne possédait plus de représentation de Sylvie Fabre, il s'épuisait à vouloir la décrire toujours plus exactement : au milieu de la cuisine naquirent des hologrammes que dégonflait la moindre imprécision. Ça ne se rend pas, soupirait Fabre en posant une main sur sa tête, sur ses yeux, et le découragement l'endormait. Souvent ce fut à Paul de déplier le canapé convertible, transformant les choses en chambre à coucher.

Description et modélisation des chaînes de références

Le projet ANR Democrat (2016-2020) et ses avancées à mi-parcours

Frédéric Landragin, Marine Delaborde, Yoann Dupont, Loïc Grobol et le consortium *Democrat*

Le projet ANR Democrat vise à développer les recherches sur la langue et la structuration textuelle du français via l'analyse détaillée et contrastive des chaînes de références (instanciations successives d'une même entité) dans un corpus diachronique de textes écrits entre le 9ème et le 21ème siècle, avec des genres textuels variés. Il réunit des chercheurs issus des laboratoires Lattice, LiLPa, ICAR et IHRIM. Il a été lancé en mars 2016 et l'essentiel des efforts porte actuellement sur l'annotation (manuelle) d'un corpus. Plusieurs expérimentations d'annotation ont eu lieu, de manière à tester différentes procédures. La procédure retenue alterne des phases manuelles et des phases automatiques pour compléter les annotations, via le lancement de scripts.

Objectifs et livrables du projet

Proposer un modèle de la référence et de la composition des chaînes de références
- modèle orienté sur le discours et pas seulement la phrase
- modèle qui s'enrichit de comparaisons inter-langues et d'études diachroniques
- perspectives : étude des transitions référentielles et de la saillance référentielle

Fournir un corpus annoté qui serve de corpus de référence et d'apprentissage
- taille visée : 500.000 mots ; 200 à 300.000 maillons de chaîne annotés
- proposer un pendant au seul corpus similaire existant pour le français : ANCOR

Développer un outil d'annotation adapté aux chaînes de références
- prototype de départ : ANALEC
- intégration des fonctionnalités d'annotation et de gestion de schémas d'annotation dans TXM

Développer un système de résolution automatique de la coréférence
- techniques d'apprentissage artificiel appliquées sur le corpus annoté manuellement
- participation envisagée à une campagne d'évaluation internationale

Publications à mi-parcours

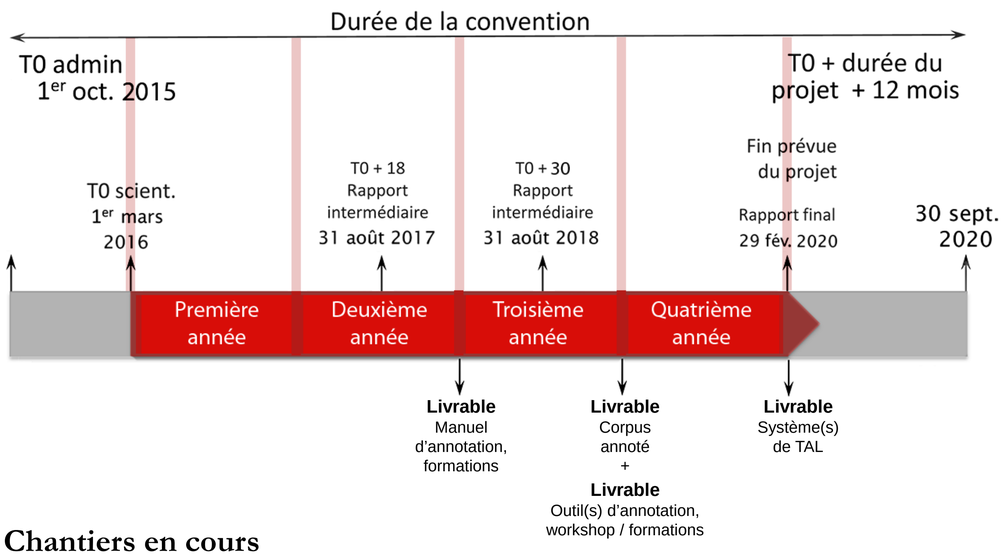
Modèle :
- une dizaine de publications, dont n° de *Langue Française*
- conférences LPTS, GLAD…

Corpus :
- articles de méthodologie
- format XML TEI

Outil d'annotation :
- TXM, ANALEC, SACR
- nouvelles métaphores IHM

Système de TAL :
- apprentissage artificiel, avec plusieurs techniques testées en parallèle (*deep learning*…)
- une dizaine de publications : TALN, TAL, CICLING, LREC…

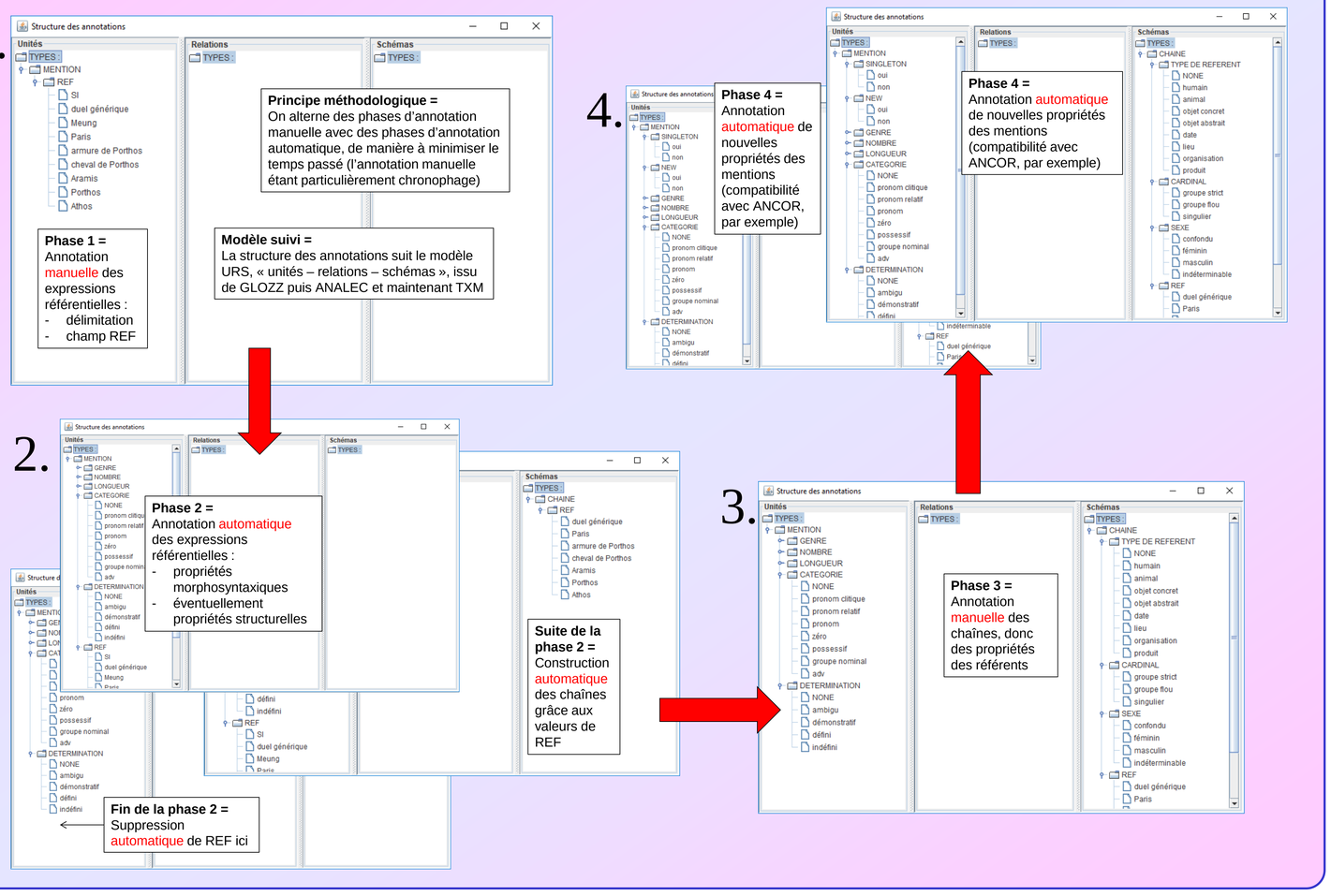
Echéances et enjeux à mi-parcours



Chantiers en cours
- liens entre chaînes de référence et structures textuelles
- approches contrastives pour l'étude cross-linguistique des chaînes de référence
- méthodologie de l'annotation manuelle pour les objets que sont les chaînes de référence
- ergonomie de l'IHM d'annotation pour la gestion d'unités, de relations et de schémas
- identification de mesures pour quantifier les analyses de chaînes et adapter au projet les possibilités d'interrogation de corpus

Enjeux pour 2018-2019
- finaliser l'annotation du corpus Democrat
- définir une procédure consensuelle d'analyse des chaînes de référence, qui serve à tous les participants et puisse faire office de procédure standard facilitant les comparaisons
- adapter les développements réalisés sur le corpus ANCOR pour le corpus Democrat

Procédure d'annotation



Références

Désoyer, A., Landragin, F., Tellier, I., Lefeuvre, A., Antoine, J.-Y. (2014) « Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR », *TAL*, 55(2), pp. 97-121.
Landragin F. (2016) « Conception d'un outil de visualisation et d'exploration de chaînes de coréférences », *Thirteen International Conference on Statistical Analysis of Textual Data (JADT 2016)*, Nice.
Landragin F., Poibeau T., Victorri B. (2012) "ANALEC: a New Tool for the Dynamic Annotation of Textual Data", *LREC 2012*, Istanbul, Turquey.
http://www.agence-nationale-recherche.fr/?Projet=ANR-15-CE38-0008
http://www.lattice.cnrs.fr/democrat/

Travail réalisé avec le soutien de l'ANR dans le cadre du projet *Democrat* ANR-15-CE38-0008