PROPOSITION ACTION POUR LA FEDERATION LI-LIFO

1- Titre Techniques d'apprentissage pour le TALN

2- Description succincte (thématiques et objectifs)

L'équipe CA du LIFO développe des recherches sur des techniques d'apprentissage automatique qui sont de plus en plus utilisées en TAL. De son côté, certains chercheurs en TAL de l'équipe BDTLN s'intéressent sans exclusive également à l'utilisation de ces techniques, sans en faire un axe de recherche spécifique. Les chercheurs en fouille de données et classification menées au sein de l'équipe BDTLN du LI ont par ailleurs également travaillé sur des données textuelles (thèse en fouille de texte de Damien Nouvel, collaboration actuelle avec l'entreprise Eloquenz sur la détection d'auteurs sur blogs). Une action LI/LIFO permettait d'approfondir ces recherches sur les techniques de TAL centrées sur les données, et ce d'autant plus qu'un des domaines d'excellence de l'équipe BDTLN réside dans la constitution de corpus qui servent précisément de données d'apprentissage dans le domaine.

Du point de vue du LIFO, l'intérêt d'un travail avec le LI sur le sujet réside avant tout sur l'étude de l'adaptation des modèles appris à un autre type de données (passage du langage écrit au langage oral dans notre cas) et de réfléchir aux liens entre traits d'apprentissage pertinents pour la linguistique et ceux qui font sens pour les techniques de classification.

Plusieurs domaines applicatifs développés dans nos laboratoires peuvent être concernés par cette thématique : détection des entités nommées, détection d'auteurs, résolution de la coréférence. Cette action se focalisera spécifiquement sur cette dernière thématique en 2015.

3- Participants (personnes impliquées avec précision sur appartenance)

Yannick Parmentier (LIFO-CA) ???????
Denys Duchier (LIFO-CA)

Anaïs Lefeuvre (LI)
Jean-Yves Antoine (LI)
Agata Savary (LI)
Nicolas Labroche (LI) ???????

4- Historique des collaborations (s'il existe des collaborations passées et des résultats déjà obtenus)

Pas de collaboration à ce jour entre le LI et le LIFO sur ce sujet spécifique. Collaboration par contre avec le laboratoire LaTTiCe (Isabelle Tellier, Frédéric Landragain).

[Desoyer 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, Traitement Automatique des Langues, *TAL*, vol. 55.

5- Proposition de travail (description plus détaillée des collaborations envisagées et des résultats attendus)

Le travail prévu dans le cadre de cette action pour 2015 se focalise donc sur la résolution des coréférences. En collaboration avec le laboratoire LLL (porteurs : Jean-Yves ANTOINE pour le LI, Emmanuel SCHANG pour le LLL), nous avons constitué dans le cadre du projet régional ANCOR le plus grand corpus de parole transcrite annoté en relations de coréférence. Ce corpus a

déjà servi de base d'apprentissage au laboratoire LaTTiCe (Isabelle TELLIER, Frédéric LANDRAGIN) pour l'apprentissage d'un petit système de résolution à base de classifieurs multiples (CROC) réalisé en utilisant la plate-forme Weka. Weka est une plate-forme générique intégrant un très grand nombre de techniques d'apprentissage pour la classification. Dans le cadre de cette action, nous souhaiterions utiliser une plate-forme dédiée spécifiquement à la résolution des co-référence (BART), qui intègre également plusieurs modèles d'apprentissage et que le LI est en train d'adapter au français. L'idée est ensuite de comparer les résultats obtenus à ceux du système CROC.

D'un point de vue scientifique, cette action revêt plusieurs objectifs :

- Comparaison et optimisation des différentes techniques de classification proposées par BART,
- Etude des relations entre traits linguistiques observés en corpus et traits utiles à l'apprentissage,
- Etude éventuelle des questions d'adaptation de modèles d'un registre oral à un autre (interviews présentes dans le sous-corpus ANCOR-ESLO et parole très interactive présente dans les sous-corpus ANCOR-OTG et ANCOR-UBS).

D'un point de vue budgétaire, cette action repose sur une demande de financement de stage de Master. Le système CROC a été précisément développé dans le même cadre (stage de Master d'Adèle Desoyer) et sur un sujet très proche, ce qui un indicateur de réussite de cette action.

6- **Prospectives** (Interactions possibles avec autre action) ? Auriez-vous des idées sur les thématiques d'un axe qui pourraient englober cette action ?

Deux perspectives de recherche à plus long terme sont clairement visées par cette action : le rapprochement de nos recherches en classification et plus généralement en apprentissage. Dans le cadre d'une application au TAL, nous avons déjà identifié les domaines applicatifs suivants

- Détection des entités nommées
- Détection d'auteur

Dans le cas du travail spécifiquement prévu sur la résolution des coréférences, la question du dépôt d'un projet ANR sur la question en collaborations avec le LaTTiCe est clairement envisageable. Ceci dépendra de la réussite du LaTTiCe d'un dépôt de projet proche cette année mais avec d'autres partenaires (le LI est associé à cette demande comme sous-traitant sur la partie Corpus) : projet DEMOCRAT actuellement en seconde phase d'évaluation.