PROPOSITION ACTION POUR LA FEDERATION LI-LIFO

1- Titre Analyse syntaxique (*parsing*) pour l'annotation de la temporalité

2- Description succincte (thématiques et objectifs)

Cette action a pour objectif de poursuivre un travail déjà entamé sur la constitution d'un corpus de parole transcrite enrichi par une annotation en arbres syntaxiques. Ceci dans la perspective d'une annotation ultérieure en temporalité, pour laquelle nous avons besoin d'un tel corpus arboré (*treebank*). Elle mobilisera pour cela l'expérience du LI (équipe BDTLN) en matière d'analyse et de représentation de la temporalité, ainsi que celle du LIFO (équipe CA) en matière de parsing syntaxique.

Les thématiques de recherche proposées ici font actuellement l'objet deux soumissions de projet :

- Dépôt d'une demande de post-doctorat centrée sur l'annotation en temporalité dans le cadre de l'APR-IA de la région centre (Jean-Yves ANTOINE – LI)
- Participation du LIFO (et à un degré moindre du LI) à une demande de financement de projet portée par le LLL dans le cadre de l'APR-IA. Cette demande est liée au corpus ESLO et comportera une sous-tâche dédiée à l'annotation en arbres syntaxiques du corpus. Le LI interviendrait précisément sur la question de l'utilisation en temporalité. Porteurs : Yannick PARMENTIER pour le LIFO, Jean-Yves ANTOINE pour le LI.

Cette proposition d'action a pour objet de formaliser ces collaborations. Dans le cas où l'un des projets mentionnés ci-dessus serait accepté, l'action proposée à la fédération ne ferait plus l'objet de demande de financement, mais plutôt de labélisation.

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)
Denis Maurel (LI)
Jakub Waszczuk (LI – doctorant)

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

Ce travail a été initié dans le cadre du projet TEMPORAL, une action de recherche soutenu par le MSH Val de Loire sur l'annotation de la temporalité. Ce petit projet collaboratif a regroupé dans un premier temps des chercheurs du LI et du LLL. Il a été porté par Emmanuel Schang (LLL) et Jean-Yves Antoine (LI) puis piloté scientifiquement par Anais Lefeuvre (LI). Les travaux des différents participants à l'action ont consisté à réfléchir à l'adaptation d'une norme ISO de représentation de la temportalité, TimeML, à la parole transcrite. Elle a conduit à l'issue du projet à la proposition d'un enrichissement de la norme TimeML consistant principalement à annoter la temporalité non plus au niveau des mots (unité lexicales) mais des nœuds d'arbres syntaxiques construits sur le corpus. Cette proposition a été validée par une communication :

- Lefeuvre A., Antoine J.-Y., Savary A., Schang E., Abouda L., Maurel D., Eskhol I.

(2014) Annotation de la temporalité en corpus : contribution à l'amélioration de la norme TimeML, *Actes TALN'2014*, Marseille

Anaïs Lefeuvre et Jean-Yves Antoine ont par ailleurs rejoint le groupe de normalisation de l'AFNOR miroir du groupe ISO TC37/SC4 pour précisément porter cette demande de modification au niveau international.

Cette évolution pose la question de la constitution de corpus arborés de parole transcrite. Le LIFO et le LI travaillant précisant de concert sur la question du parsing (thèse de Jakub Waszczuk co-encadrée par Yannick Parmentier pour le LI et Agata Savary pour le LIFO ; pilotage par les mêmes encadrants de l'action COST PARSEME), les laboratoires LI, LIFO et LLL ont décidé de poursuivre la collaboration initiée par TEMPORAL sur la temporalité en y intégrant la question de l'annotation syntaxique nécessité par notre proposition de format de représentation alternatif à TimeML. Le LIFO a ainsi déjà participé à une réunion scientifique post-TEMPORAL

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

Comme préciser plus haut, cette action de recherche fait l'objet de deux dépôts de demande de financement sur les deux problématiques scientifiques que soulève cette collaboration : l'annotation syntaxique de corpus de parole transcrite d'une part, et son utilisation pour la représentation de la temporalité d'autre part.

L'action que nous proposons dans le cadre de la fédération a pour objectif d'assurer la réalisation d'une avancée minimale mais absolument nécessaire à la poursuite de la réflexion dans le cas d'une non-acceptation sur les appels à projets précités : l'annotation en arbre syntaxique de corpus de parole transcrite.

Elle consistera donc au financement d'un stage de Master centré sur l'adaptation au langage oral de parseurs syntaxiques ainsi qu'à la production d'un petit corpus pilote arboré grâce à ces outils (annotation automatique puis révision manuelle). Cette étude sera menée au choix sur les parseurs développés par le LIFO ou sur l'adaptation de parseurs disponibles en open source.

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 :

- Annotation de la temporalité avec un paradigme de représentation enrichi
- Constitution d'un large corpus de parole annoté en arbre syntaxique

Cette action sur l'analyse syntaxique entre en interaction avec l'action « Expressions polylexicales et parsing » portée par les équipes BDTLN du LI et CA du LIFO.