

Opération 7

Traitement automatique des langues (TAL)

Composition

Responsable : L. Tanguy

Chercheurs et enseignants-chercheurs : D. Bourigault [& Op. 3, 5] ; C. Fabre [& Op. 5] ; N. Hathout [& Op. 2, 5] ; M.-P. Péry-Woodley [& Op. 4, 5] ; J. Rebeyrolle [& Op. 4, 5] ; L. Tanguy [& Op. 5]

I.T.A. : F. Sajous [& Op. 5]

Doctorants : J. Eychenne [& Op. 1] ; C. Frérot [& Op. 5] ; E. Galy ; M. Ho-Dac [& Op. 4, 5] ; A. Josselin-Leray [& Op. 5] ; M. Laignelet [& Op. 4, 5] ; S. Ozdowska [& Op. 5] ; C. Pimm [& Op. 4, 5]

Post-doctorants : M.-P. Jacques [& Op. 5]

1. État des lieux des activités de TAL à l'ERSS au cours de la période 2003-2006

1.1. Activités liées au TAL au sein de l'ERSS

Au début de la période correspondant au plan quadriennal 2003-2006, les activités de TAL ont été menées exclusivement au sein de l'opération *Sémantique et Corpus*, autour du traitement de corpus et de l'analyse terminologique. Par la suite, du fait en particulier de la part croissante du travail sur corpus électronique au sein de l'ERSS, des activités liées au traitement automatique ont été entreprises dans les différents domaines de la linguistique couverts à l'ERSS.

- **phonologie** : repérage automatisé de phénomènes phonologiques (élision, schwa), annotation morphosyntaxique de corpus d'oral retranscrit.
- **morphologie** : construction de ressources morphologiques pour le TAL, création d'outils d'investigation (Webaffix), études quantitatives des données morphologiques en corpus.
- **syntaxe** : développement d'un analyseur syntaxique (Syntex), environnement de recherche dans des corpus analysés, adaptation à l'analyse syntaxique de l'oral retranscrit, développement de ressources linguistiques pour l'analyse syntaxique (données de sous-catégorisation).
- **lexique** : extraction de termes, analyse distributionnelle automatique (Upery), désambiguïsation lexicale, extraction de marqueurs de relation (Yakwa).
- **discours** : mise au point et projection de marqueurs pour la segmentation discursive.

Ces activités ancrées dans les domaines de recherche des diverses opérations sont complétées par plusieurs activités transversales à l'Equipe :

- mise au point d'une chaîne d'annotation de corpus : segmentation, étiquetage, analyse syntaxique, formatage XML
- activités de formation (aux outils d'analyse de corpus, à XML)
- projet ARIEL pour la recherche d'information (traitements morphologiques, syntaxiques, sémantiques)

1.2. Spécificités du TAL dans une équipe de linguistique

Les activités dans le domaine du Traitement Automatique des Langues à l'ERSS se sont donc diversifiées, mais elles présentent le caractère commun de s'articuler de façon forte avec des recherches en linguistique. Ces activités s'organisent principalement selon deux axes :

- mise en place d'environnements pour l'observation linguistique (de la phonologie au discours)
- création de ressources linguistiques et d'outils pour le Traitement Automatique des Langues et pour ses applications

1.3. Première étape de structuration : l'axe TAL

Pour mieux prendre en compte ces évolutions des activités TAL au sein de l'ERSS, un axe TAL est créé en décembre 2004. Il a regroupé 19 personnes, dont 9 permanents. Des groupes de travail se sont constitués en son sein et ont abordé les thèmes suivants :

- chaîne d'étiquetage de corpus écrits : créer une version commune et optimale de l'étiqueteur utilisé - Treetagger
- outils d'interrogation de corpus : partager les compétences sur l'utilisation d'outils créés ou utilisés au sein de l'équipe
- technologies XML et XSLT : utiliser les différentes modalités d'indexation de corpus au format XML
- statistiques pour l'analyse linguistique : améliorer et étendre le recours aux méthodes quantitatives
- inventaire des ressources textuelles et lexicales constituées ou utilisées par les membres de l'équipe

2. Création d'une opération TAL

En Juin 2005, l'opération TAL a été officiellement créée au sein de l'ERSS. Elle remplace alors l'axe transversal TAL, dont elle reprend les activités mais en étendant les problématiques scientifiques.

2.1. Rapprochement des domaines d'analyse

La création d'une opération TAL est motivée d'abord par la volonté d'aller au-delà du partage de techniques et de compétences transversales, pour mettre en place des projets articulant les travaux réalisés jusqu'ici en parallèle autour du TAL, sous l'égide de différentes sous-disciplines de la linguistique. Il s'agit donc de mettre en place des chantiers réunissant plusieurs domaines d'activité, par exemple la phonologie et la syntaxe, avec le développement d'une chaîne d'annotation syntaxique automatique de l'oral retranscrit, ou le discours et le lexique, avec l'utilisation d'indices lexicaux pour la segmentation discursive automatique.

2.2. Partage des compétences sur le TAL

La diversification des activités TAL au sein de l'ERSS exige un lieu pour une mise en commun des compétences et expériences, et pour des échanges sur les divers aspects de la recherche en TAL : épistémologiques (quel est le retour des applications sur l'appréhension des problématiques linguistiques, quel est le rôle de la linguistique dans le développement d'outils de TAL), méthodologiques (veille scientifique, prise en compte des nouvelles approches) et techniques (mise au point et utilisation d'outils, continuité des activités de l'axe TAL).

2.3. Partenariat, affichage

Avec la mise en place d'une opération TAL, l'ERSS améliore sa visibilité sur ce domaine de recherche, ce qui lui permet, d'une part, de se constituer en partenaire pour des collaborations au sein de la

communauté informatique et TAL, au niveau régional (renforcement d'une collaboration avec l'Institut de Recherche en Informatique de Toulouse), national (réponse à des appels à projet CNRS ou ministère de la recherche) et international (projets européens), et, d'autre part, de construire un contexte de recherche attractif pour les étudiants de troisième cycle, dans la continuité de la filière TAL mise en place au département sciences du langage (parcours TAL du master recherche).

3. Prospective de l'opération TAL

3.1. Une problématique fédératrice : le passage à l'échelle

Au travers des activités menées en son sein, en collaboration ou non avec d'autres opérations, l'opération TAL se concentrera sur le thème fédérateur du *passage à l'échelle* auquel est confronté le travail linguistique sur données attestées. A ce titre l'utilisation du Web comme corpus linguistique est emblématique. Les différents pans de la linguistique connaissent depuis quelques années une montée en puissance de l'utilisation de données, et ces données sont de plus en plus volumineuses. Les phénomènes d'échelle se situent à différents niveaux : pour la morphologie et les études sur le lexique, la récolte de nouvelles attestations permet un regard nouveau sur des phénomènes jusqu'ici trop éparés pour avoir pu être étudiés. Pour la syntaxe, la disponibilité d'analyseurs syntaxiques robustes permet d'envisager la mise en place d'un observatoire syntaxique du français sur des corpus divers et volumineux. Pour le lexique, la multiplicité des ressources textuelles permet d'approcher les problèmes de la variation contextuelles et de la caractérisation des usages. Pour les études sur le discours, la simple prise en compte de corpus est déjà une approche nouvelle, mais nécessite des outils spécifiques pour l'annotation et l'exploitation de ces données.

Ce passage à l'échelle a également des implications sur le plan méthodologique pour le TAL :

- sur le plan des données manipulées, la croissance exponentielle est due à la numérisation de corpus traditionnels, et surtout à l'exploitation du Web. La transition vers l'utilisation du Web comme (source de) corpus doit être accompagnée d'une réflexion méthodologique plus poussée sur la caractérisation des données (collaboration avec l'opération 5). A ces considérations s'ajoutent bien entendu le besoin d'un outillage spécifique pour l'accès à ces données et leur annotation. Le projet principal concernera donc la mise en place d'un outil de parcours du Web à visée linguistique, dont les résultats et le paramétrage dépendront des disciplines impliquées : la morphologie pour le repérage de nouveaux dérivés, la syntaxe pour la constitution de corpus annotés (voir point suivant), la caractérisation des corpus en genres, etc.

- sur le plan des annotations et de l'exploitation, les points principaux concerneront la création d'un observatoire de la langue à partir de l'annotation syntaxique par Syntex d'un corpus diversifié, l'annotation discursive, le profilage de documents pour la caractérisation des genres. D'un point de vue technique, ces différents travaux porteront également sur les normes d'annotation et l'utilisation des technologies XML.

- sur le plan des applications, la disponibilité de données massives permet également au TAL de proposer des ressources pertinentes pour des applications en ingénierie linguistique. Le principe du passage à l'échelle est également une étape dans la validation des techniques du TAL quand celles-ci sont confrontées à des environnements applicatifs en vraie grandeur, comme c'est notamment le cas pour la recherche d'information, discipline consommatrice de ressources linguistiques robustes pour des applications manipulant des données volumineuses et très variées.

3.2. Activités et projets

Les chantiers qu'il est prévu de mettre en place sont les suivants :

Analyse syntaxique

- analyse de la subordination dans le cadre de l'analyseur Syntex
- traitement de nouvelles langues (espagnol et basque)

- mise en place d'un observatoire syntaxique du français : proposition d'un serveur central pour l'analyse et l'exploitation de corpus annotés syntaxiquement, utilisable par l'ensemble de la communauté scientifique

Annotation et exploitation de corpus

- étude des modalités d'annotation de phénomènes discursifs, en plus des informations morphosyntaxiques et syntaxiques

Ressources pour le TAL

- développement de bases de données lexicales génériques, et notamment de classes sémantiques acquises à partir de corpus ou de dictionnaires de langue
- extension des bases de données morphologiques, et mise à disposition pour la communauté

Exploitation du Web

- acquisition de créations lexicales, via le déploiement d'un « crawler » dédié à cette tâche

TAL et Recherche d'information

- évaluation des ressources linguistiques dans le cadre de la recherche d'information (bases morphologiques et sémantiques du français)

Approches statistiques

- caractérisation des genres par des marqueurs linguistiques
- étude des titres.