Préface

Le développement de systèmes de TAL, que ce soit en recherche ou en contexte applicatif, fait de plus en plus fréquemment appel à des infrastructures logicielles complexes. Plusieurs facteurs expliquent ce mouvement : travaux sur corpus, souvent de grande taille et de formats diversifiés, nécessitant un cycle modélisation–expérimentation–évaluation court ; articulation de traitements et de ressources linguistiques de natures et de provenances diverses ; réutilisation et capitalisation de ces modules et de ces ressources ; sophistication des modèles linguistiques mis en œuvre… Tous ces facteurs militent de concert pour la réalisation et l'usage de véritables *environnements de développement* dédiés au TAL. L'usage de ces environnements permet d'importants gains de productivité dans la réalisation de systèmes opérationnels aussi bien que dans la mise au point des modèles linguistiques eux-mêmes.

Un nombre important de telles *plates-formes pour le TAL* a vu le jour ces dernières années, tant en milieu industriel qu'académique, ou au sein de projets collaboratifs. Si quelques « majors » se détachent par leur ancienneté et l'étendue du cercle de leurs utilisateurs (GATE, UIMA, Open NLP, Nooj, Unitex…), on observe en fait une remarquable floraison de systèmes dont le présent volume se fait l'écho. Cette « biodiversité » n'est l'effet ni du hasard ni de la propension naturelle des informaticiens à toujours réécrire de nouveaux systèmes, mais résulte de la prise en compte de divers types de contraintes de nature à la fois informatique et linguistique dépendant des tâches visées et des conditions de déploiement, conduisant à des choix complexes et articulés :

– choix d'une architecture logicielle pour combiner les différents traitements : pipeline, agents, services Web distribués… ;

– représentation des documents et des annotations apportées par les modules successifs : robustesse à la variabilité des formats, prise en compte du multilinguisme ou de la multimodalité… ;

– variété des ressources linguistiques intégrées dans la plate-forme (modèles d'analyse, grammaires, lexiques…), ouverture éventuelle du système à des ressources externes ;

– interfaces permettant aux utilisateurs d'assembler des « briques » de traitement et de visualiser leur projection en corpus ;

– etc.

L'ensemble de ces questions est débattu dans les contributions recueillies pour le présent volume. Dix plates-formes différentes sont présentées, les auteurs s'attachant à en décrire les fonctionnalités, les objectifs et les principes de conception, ainsi que des applications ou expérimentations réalisées grâce à elles.

Un premier groupe de trois articles présentent des *plates-formes généralistes*, ayant vocation à articuler une grande variété de traitements pour des tâches *a priori* quelconques. Elles sont très significatives de ce que l'on peut appeler le *mainstream* du domaine :

1. Johannes Heinecke, Grégory Smits, Christine Chardenon, Emilie Guimier De Neef, Estelle Maillebuau et Malek Boualem présentent *TiLT : plate-forme pour le traitement automatique des langues naturelles* développée au sein d'un laboratoire industriel, Orange Labs. Cette plate-forme intègre une importante panoplie de ressources et de modules d'analyse : correction lexicale, morphologie, analyse syntaxique (*chunking* et dépendances), analyse sémantique (réseaux sémantiques) avec une attention particulière sur la prise en compte du multilinguisme. Un modèle de décision multicritère permettant la gestion d'hypothèses d'analyses concurrentes est proposé et l'article présente également plusieurs applications « opérationnelles » impliquant de fortes contraintes en termes de robustesse et de portabilité.

2. Dans *Antelope. Une plate-forme industrielle de traitement linguistique* François-Régis Chaumartin, prenant acte du développement de ressources et d'analyseurs disponibles dans la communauté scientifique, présente une architecture permettant de l'intégration de composants externes au sein d'une plate-forme qui demeure néanmoins « homogène ». L'accent est mis sur le respect de principes généraux de génie logiciel. Sur le plan linguistique, Antelope repose sur le modèle Sens-Texte de Mel-čuk et donne notamment accès de manière unifiée à diverses ressources sémantiques lexicales « autour de WordNet ». La plate-forme a elle aussi été développée en contexte industriel (société Proxem) et plusieurs applications sont évoquées.

3. Avec *Articulation des traitements en TAL. Principes méthodologiques et mise en œuvre dans la plate-forme LinguaStream*, Antoine Widlöcher et Frédérik Bilhaut proposent une réflexion méthodologique sur les conditions de la bonne articulation, au sein d'une plate-forme de TAL, des traitements hétérogènes requis par les travaux sur corpus. Un ensemble de principes linguistiques et informatiques sont explicités : approche logicielle « par composants » ; prise en compte de la diversité des objets et structures linguistiques manipulés, impliquant la coexistence de divers modèles génériques d'analyses (grammaires et transducteurs, ressources lexicales, outils lexicométriques…) ; définition d'un modèle d'annotation unifiée ; outils d'observation des corpus analysés… L'article montre comment ces principes sont mis en œuvre dans la plate-forme LinguaStream dont trois applications sont présentées et discutées.

Trois autres articles s'intéressent à l'utilisation de plates-formes pour mettre en œuvre des *applications* ou *des tâches spécifiques* :

4. Horacio Saggion, *SUMA, A Robust and Adaptable Summarisation Tool*, présente un système générique pour le développement et l'évaluation de logiciels de résumé automatique réalisé dans le cadre de la plate-forme GATE. Le système bénéficie de l'environnement GATE pour un certain nombre de tâches (stockage,

enchaînement des traitements, annotation des textes, visualisation…). L'article décrit un ensemble de ressources spécifiques pour le résumé constituant le « *SUMMA Toolkit* », offrant en particulier des fonctionnalités multilingues et multidocuments.

5. Le travail présenté par Thierry Hamon et Adeline Nazarenko dans *Le développement d'une plate-forme pour l'annotation spécialisée de documents Web : retour d'expérience*, concerne plus particulièrement la réalisation de moteurs de recherche « sémantiques » pour des domaines spécialisés. Les auteurs ont développé une plate-forme, Ogmios, qui intègre des modules classiques en recherche et extraction d'information, de la reconnaissance d'entités nommées à l'étiquetage sémantique en passant par la résolution d'anaphores. Une architecture distribuée permet d'accélérer les traitements. Plusieurs expériences d'utilisation sont présentées et évaluées, conduisant à un ensemble de réflexions méthodologiques sur des questions telles que l'évaluation des plates-formes, les outils d'expérimentation sur corpus, l'articulation local/global, ou la sémantique des annotations.

6. Dans *SxPipe 2 : architecture pour le traitement présyntaxique de corpus bruts*, Benoît Sagot et Pierre Boullier posent la question des prétraitements (dits encore « traitements de surface ») nécessaires à l'exploitation de corpus « bruts » en amont de diverses tâches de TAL. Ils présentent une plate-forme générique et configurable, SxPipe 2 qui décompose ces traitements en un certain nombre d'étapes, réalisés par autant de modules spécifiques interchangeables. Une représentation en DAG du texte analysé permet de gérer les ambiguïtés. Un dispositif original de reconnaissance de motifs est proposé, ainsi que divers composants réalisant des tâches courantes telles que la correction orthographique, la reconnaissance d'entités nommées et de mots composés, le découpage en tokens ou la segmentation en phrases.

Deux articles concernent la question clé des *annotations* apportées au texte pour représenter les informations linguistiques qui en sont extraites, et plus spécifiquement la *fusion d'annotations* de provenances diverses :

7. Sylvain Loiseau, dans *CorpusReader : construction et interrogation de corpus multiannotés*, pose la question de l'articulation des traitements sous un jour différent des plates-formes *mainstream*. Prenant acte de la multiplication d'outils produisant des annotations de différents niveaux (morphologique, syntaxique, sémantique…), il propose d'opérer leur fusion *a posteriori*, plutôt que leur production par plusieurs modules au sein d'un unique environnement. Il argumente par ailleurs sur l'intérêt, pour l'observation linguistique, de la combinaison de ces différents niveaux, susceptible de faire apparaître de nouvelles corrélations pertinentes. La plate-forme *CorpusReader* permet d'opérer la fusion d'annotations et l'exploitation des corpus multiannotés pour en extraire des sous-corpus, des représentations ou des quantifications.

8. Dans *A flexible Framework for Integrating Annotations from Different Tools and Tagsets*, Christian Chiarcos, Stefanie Dipper, Michael Götze, Ulf Leser,

Anke Lüdeling, Julia Ritz, Manfred Stede posent également le problème de l'exploitation d'annotations provenant de diverses sources (avec une motivation particulière pour l'annotation manuelle et les langues faiblement représentées). Leur contribution est focalisée sur la question de l'interopérabilité des formats d'annotation : ils proposent un format pivot, PAULA, ainsi qu'une ontologie des annotations linguistiques permettant de représenter la correspondance entre différents jeux d'étiquettes. La plate-forme ANNIS permet de gérer ces correspondances en réalisant les conversions nécessaires et ainsi d'effectuer des recherches sur des corpus relevant de modèles d'annotation hétérogènes.

Le processus *d'évaluation* peut aussi tirer bénéfice de plates-formes spécifiques :

9. C'est ce que mettent en évidence Olivier Hamon, Patrick Paroubek et Djamel Mostefa dans *SEWS : un serveur d'évaluation orienté Web pour la syntaxe*. Plus précisément, il s'agit d'une plate-forme dédiée à l'évaluation des systèmes d'analyse syntaxique, développée et expérimentée dans le cadre de la campagne d'évaluation PASSAGE. Les auteurs en présentent les principales fonctionnalités et l'architecture logicielle ; ils discutent le problème de la compatibilité des différents formats d'annotation et décrivent les principes du format EASY retenu. Le compte-rendu d'expérience montre les gains d'efficacité de différents ordres que permet une telle plate-forme, tant pour les participants que pour les organisateurs.

Enfin, un des articles pose de manière en quelque sorte radicale la question de *l'implémentation* des plates-formes de TAL :

10. Dans *Cocytus : parallel NLP over disparate data*, Noah Evans, Masayuki Asahara et Yuji Matsumoto observent que les systèmes d'exploitation de type Unix offrent déjà un ensemble de dispositifs pour enchaîner des traitements informatiques et gérer des flux de données. Ils défendent alors l'idée selon laquelle des développements appropriés peuvent en faire une alternative efficace aux plates-formes « logicielles ». Ils proposent le système Cocytus, basé sur le système d'exploitation Inferno. Un modèle de représentation de structures arborescentes, adapté au traitement en flux, est défini pour coder des données linguistiques. De plus une accélération des traitements peut être obtenue grâce à une parallélisation transparente pour l'utilisateur. Une évaluation de performances sur le *Penn Treebank* est présentée.

On le voit, sans prétendre à l'exhaustivité, l'ensemble de ces contributions témoigne de la richesse d'un domaine de recherche que nous pensons prometteur et source d'importants gains de productivité dans le développement des applications TAL.

Patrice Enjalbert
Laboratoire GREYC (UMR 6072)
UFR Sciences - Université de Caen - Campus II
Bd. Maréchal Juin - B.P. 5186 14032 C aen Cedex
Patrice.Enjalbert@info.unicaen.fr