ALPAGE - Analyse linguistique profonde à grande échelle

Rapport Hcéres

agence d'évaluation de la recherche
et de l'enseignement supérieur

Section des Unités de recherche

Evaluation de l'AERES sur l'unité :

Analyse Linguistique à Grande Echelle

ALPAGE

sous tutelle des établissements et organismes :

Université Paris 7 - Denis Diderot

Institut National de Recherche en Informatique et en Automatique

Décembre 2012

agence d'évaluation de la recherche
et de l'enseignement supérieur

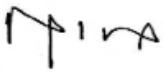
Section des Unités de recherche

Le Président de l'AERES

Didier Houssin

Section des Unités de recherche

Le Directeur



Pierre Glaudes

Le président de l'AERES "signe [...], les rapports d'évaluation, [...] contresignés pour chaque section par le directeur concerné" (Article 9, alinéa 3, du décret n°2006-1334 du 3 novembre 2006, modifié).

Notation

À l'issue des visites de la campagne d'évaluation 2012-2013, les présidents des comités d'experts, réunis par groupes disciplinaires, ont procédé à la notation des unités de recherche relevant de leur groupe (et, le cas échéant, des équipes internes de ces unités). Cette notation (A+, A, B, C) a porté sur chacun des six critères définis par l'AERES.

NN (non noté) associé à un critère indique que celui-ci est sans objet pour le cas particulier de cette unité ou de cette équipe.

Critère 1 - C1 : Production et qualité scientifiques ;
Critère 2 - C2 : Rayonnement et attractivité académique ;
Critère 3 - C3 : Interaction avec l'environnement social, économique et culturel ;
Critère 4 - C4 : Organisation et vie de l'unité (ou de l'équipe) ;
Critère 5 - C5 : Implication dans la formation par la recherche ;
Critère 6 - C6 : Stratégie et projet à cinq ans.

Dans le cadre de cette notation, l'unité de recherche concernée par ce rapport a obtenu les notes suivantes.

- Notation de l'unité : Analyse Linguistique Profonde À Grande Échelle - ALPAGE

C1	C2	C3	C4	C5	C6
A	A	A+	A+	A+	A

Rapport d'évaluation

Nom de l'unité :	Analyse Linguistique à Grande Echelle
Acronyme de l'unité :	ALPAGE
Label demandé :	UMR-I
N° actuel :	UMR-I 001
Nom du directeur (2012-2013) :	Mme Laurence DANLOS
Nom du porteur de projet (2014-2018) :	M. Benoît SAGOT

Membres du comité d'experts

Président : M. Patrick SAINT-DIZIER

Experts : M. Frédéric BIMBOT (Représentant les CSS de l'INRIA)

Mme Fiammetta NAMER (Représentante du CNU)

Délégué scientifique représentant de l'AERES :

M. Olivier ROUX

Représentant(s) des établissements et organismes tutelles de l'unité :

M. Richard LAGANIER, Université Paris 7 - Denis Diderot

Mme Isabelle RYL, INRIA

1 • Introduction

Historique et localisation géographique de l'unité

L'unité ALPAGE a été créée en tant qu'équipe-projet Inria en Juillet 2007 par l'Inria sur un projet en traitement automatique des langues. En Janvier 2009, elle est devenue la première UMR ; elle est associée à l'université Paris 7 Denis Diderot.

L'unité ALPAGE est située majoritairement dans les locaux de Paris 7, avec un bureau à l'Inria Rocquencourt. Elle compte 7 chercheurs permanents (2 CR INRIA, 1 PR, 3 MCF et un DR émérite), 8 doctorants, quelques postdocs et ingénieurs sous contrat et 1 assistante de projet à 50%. Seuls les PR et DR ont une HDR. Les doctorants actuels sont financés par divers moyens : allocation de recherche, bourse CIFRE ou contrats d'entreprise.

Tous les membres de l'unité sont publiants.

Équipe de Direction

La direction est assurée jusqu'à ce jour par M^me Laurence DANLOS, PR. En 2014, il est prévu que ce soit M. Benoît SAGOT, actuellement CR1 Inria, qui prenne la suite.

Nomenclature AERES

ST6 Sciences et technologies de l'information et de la communication

SHS4_1 Linguistique

Effectifs de l'unité

Effectifs de l'unité	**Nombre au 30/06/2012**	**Nombre au 01/01/2014**	2014-2018 **Nombre de produisants du projet**
N1 : Enseignants-chercheurs titulaires et assimilés	5	4	4
N2 : Chercheurs des EPST ou EPIC titulaires et assimilés	3	2	2
N3 : Autres personnels titulaires (n'ayant pas d'obligation de recherche)	1	1	1
N4 : Autres enseignants-chercheurs (PREM, ECC, etc.)			
N5 : Autres chercheurs des EPST ou EPIC (DREM, Post-doctorants, visiteurs etc.)	1	2	2
N6 : Autres personnels contractuels (n'ayant pas d'obligation de recherche)			
TOTAL N1 à N6	10	9	9

Taux de produisants	*100 %*

Effectifs de l'unité	Nombre au 30/06/2012	Nombre au 01/01/2014
Doctorants	11	
Thèses soutenues	1	
Post-doctorants ayant passé au moins 12 mois dans l'unité *	3	
Nombre d'HDR soutenues		
Personnes habilitées à diriger des recherches ou assimilées	2	2

2 • Appréciation sur l'unité

L'unité Alpage développe plusieurs thématiques de recherche en traitement automatique du langage naturel (TALN). A travers celles-ci, l'unité vise à définir une chaîne de traitements les plus complets possibles de la langue, allant de la morphologie jusqu'à la sémantique. L'unité Alpage s'est donné comme finalité de développer des outils et des ressources pour le français, à la fois pertinents linguistiquement (et pour certains d'entre eux conçus pour être reproduits dans d'autres langues) et adéquats au niveau des emplois informatiques. Ces outils/ressources sont librement distribués. Pendant les 4 années passées, l'effort a été dédié au français, mais des collaborations mises en place (ainsi que la nature générique des outils) ont rendu possible l'exploration d'autres langues, y compris pour lesquelles il y a encore peu de ressources.

Une perspective importante vise à développer une chaîne de traitements TALN écrit qui va de la morphologie jusqu'à la sémantique.

Les perspectives managériales et de direction qui sont proposées sont tout à fait pertinentes et assurent une bonne stabilité à l'unité.

Points forts et possibilités liées au contexte

Le rayonnement scientifique est de qualité, il se manifeste par des expertises, des invitations, des publications et de nombreux projets ANR en particulier.

Il y a une bonne synergie avec le monde académique, les autres laboratoires de recherche, et l'industrie. Le projet est en cohérence avec les réalisations de la période 2007-2012.

L'équipe est très soudée, ses membres sont complémentaires avec toutefois une majorité qui sont orientés vers les méthodes d'analyse statistique ou hybrides. L'intégration des étudiants est excellente.

Il y a deux volets novateurs dans le projet : l'analyse des productions spontanées (ou bruitées) des usagers de la Toile et l'étude des relations sémantique-pragmatique (dont les contours doivent toutefois être précisés, tant le domaine est vaste).

On note de très bons rapports avec Inria qui assure la gestion de l'unité, en accord avec Paris 7. L'entretien avec les tutelles a fait ressortir qu'Inria soutient pleinement cette unité qui a bien la taille et les résultats d'une équipe projet Inria.

Points à améliorer et risques liés au contexte

En ce qui concerne la linguistique descriptive et théorique, il semble que les moyens et les références soient un peu en retrait par rapport à l'état de l'art relativement aux compétences requises pour une innovation optimale dans la modélisation, la formalisation et l'implémentation.

On peut noter un foisonnement de projets : on peut se demander s'il y a des forces suffisantes pour tout mener de front ?

L'unité comporte seulement 2 HDR dont un DR émérite.

Les projets de recherche semblent en grande partie tournés vers le court ou le moyen terme au dépens de problématiques plus complexes envisagées sur le long terme.

On note, enfin, un manque de lisibilité du positionnement scientifique (et pas seulement de la place) par rapport au contexte national et international.

Recommandations

A coté d'activités bien stabilisées et pérennes, il serait bon de définir, sous la forme d'une activité éventuellement secondaire, une thématique en recherche fondamentale, avec réflexion et « prise de risque » sur les thématiques linguistiques. L'unité a en effet tout intérêt à se doter d'une visibilité de ce point de vue là, dont les étudiants, en particulier, pourront tirer bénéfice.

On peut recommander à l'unité :

- de faire passer des HDR aux chargés de recherche et aux maîtres de conférences, et aussi d'inciter davantage les étudiants à effectuer des séjours dans des laboratoires extérieurs, étrangers en particulier ;
- de réfléchir à une hiérarchisation des projets, en lien avec les priorités du laboratoire.

3 • Appréciations détaillées

Appréciation sur la production et la qualité scientifiques

La production de publications est satisfaisante (17 revues, 148 actes de colloques, 7 chapitres d'ouvrage souvent en collaboration avec des auteurs d'autres unités, assez nombreux workshops diversifiés), production de ressources libres de droits, bon positionnement dans les campagnes d'évaluation. Toutefois, il convient aussi de développer les publications en linguistique plus théorique. Il serait souhaitable de soutenir davantage l'innovation dans ce champ : les principes linguistiques théoriques sont exploités par l'unité, mais ils ont rarement été élaborés par elle.

Appréciation sur le rayonnement et l'attractivité académiques

L'unité mène ou est sur le point de démarrer plusieurs projets nationaux financés (ANR, FUI). Elle participe activement au Labex EFL. Elle collabore avec des unités de recherche en France et à l'étranger. Elle est sollicitée pour son expertise dans des jurys, comités d'évaluation, conférences, revues ; elle joue un rôle important à l'association pour le traitement des langues (ATALA). Enfin, elle constitue un pôle d'attraction pour les chercheurs (invités, post-docs).

L'unité a un excellent rayonnement national, et un bon rayonnement international.

Appréciation sur l'interaction avec l'environnement social, économique et culturel

L'unité participe à l'organisation de plusieurs manifestations (par exemple, les 50 ans de l'Atala) elle développe la vulgarisation scientifique et des synergies avec l'industrie. Plusieurs transfers technologiques vers des entreprises ont eu lieu, et il y a même eu création d'une start-up.

Appréciation sur l'organisation et la vie de l'unité

L'unité est un petit groupe bien soudé, entretenant de bons rapports et donnant à voir une bonne organisation dans le suivi des étudiants. Des séminaires réguliers et de qualité sont organisés. Les décisions sont prises de façon collégiale avec la participation de l'ensemble des membres de l'unité: c'est une forme de démocratie on ne peut plus souhaitable, qui garantit l'harmonie des relations que l'on ressent entre ces membres (permanents ou non), et qui est certainement pour beaucoup dans le dynamisme de l'unité.

Appréciation sur l'implication dans la formation par la recherche

Les membres de l'unité sont intimement liés au cursus (licence et maitrise) de Linguistique Informatique de Paris 7, qu'ils coordonnent et où ils interviennent. On note aussi plusieurs échanges d'étudiants avec des laboratoires étrangers. Enfin, toutes les thèses sont financées. De plus, elles sont toutes soutenues en moins de 4 ans. L'unité encadre des thèses en sciences du langage (ED 132) et en informatique (ED 386).

Les doctorants assistent à différents séminaires : le séminaire ALPAGE, la journée annuelle des doctorants, les séminaires organisés par des unités de recherche géographiquement proches, par exemple, celui du LLF, laboratoire de linguistique formelle de Paris Diderot.

Appréciation sur la stratégie et le projet à cinq ans

Tout d'abord, l'unité étant jeune, elle poursuit naturellement les projets démarrés en 2007, en privilégiant et approfondissant les aspects saillants (multilinguisme, convergence d'algorithmes et de ressources, …). Le but est de maîtriser l'intégralité de la chaîne de traitement de l'écrit, ce qui est très ambitieux.

En complément, l'équipe développe des projets novateurs, en particulier l'exploration du contenu spontané de la Toile « authentique », en partie pour en normaliser les aspects lexicaux, grammaticaux et syntaxiques. Par ce dernier projet, l'unité pourrait constituer un pôle d'un axe de recherche réunissant tous ceux qui s'intéressent à la Toile comme source de données, quel que soit le point de vue adopté. Une autre dimension intéressante, mais plus classique, du projet, vise à une analyse de l'interface sémantique-pragmatique, via le développement de ressources.

On peut remarquer que les projets sont plutôt de caractère applicatif, et sur des perspectives de moyen terme. Ils ne font pas ressortir la dimension fondamentale linguistique que l'on pourrait attendre.

La stratégie de l'unité est viable, mais représente une quantité de travail considérable au regard du petit nombre de membres de l'unité. Toute recherche à visée applicative en TALN est en effet très consommatrice de temps.

4 • Analyse thème par thème

Thème 1 : Morphologie et syntaxe

Effectifs : (toute l'unité a priori)

• Appréciations détaillées

Pour ce qui est du bilan, Il porte essentiellement sur l'axe de recherche le plus ancien et il engage donc le plus grand nombre de chercheurs de l'unité. Quatre objectifs ont été poursuivis pendant la période passée (2007-2012) :

1) l'optimisation de la couverture lexicale des analyseurs syntaxiques et de leur robustesse (ceci inclut : l'analyse morphologique, l'acquisition de ressources lexicales, le traitement morphologique, l'étiquetage catégoriel, et la reconnaissance d'entités nommées), ceci a donné lieu aux actions SxPipe, MElt, formatisme lexical Alexina et outils associés ;
2) les travaux en grammaire formelle, notamment sur les grammaires faiblement sensibles au contexte, les méta-grammaires (FRMG : testé avec succès lors de campagnes d'évaluation et dans le cadre de projets ANR), et le parseur hybride TAG/TIG ;
3) la mise en place d'analyseurs statistiques modernes, réunis sous la forme d'une boîte à outils (la chaîne de traitement modulaire Bonsai) ;
4) La constitution dans le formalisme Alexina du lexique Lefff pour le français et d'autres lexiques pour diverses langues.

En ce qui concerne le projet de l'unité, un axe a été fixé, qui consiste à développer en parallèle deux directions dans les travaux et recherches :

- les langages formels et la linguistique formelle ;
- la synergie dans le développement d'analyseurs syntaxiques, de manière à faire collaborer analyse symbolique et analyse statistique.

L'accent est mis sur l'analyse morpho-syntaxique des langues à morphologie riche avec deux innovations majeures :

- en syntaxe, le passage aux textes relevant d'un domaine différent de ceux que l'analyse statistique exploite traditionnellement dans son étape d'apprentissage (i.e. le contenu des journaux d'information) ;
- en morphologie lexicale et computationnelle : le passage au contenu de la Toile non contrôlée.

Pour mener à bien ce projet, les membres de ce thème :

1) mettent sur pied des collaborations avec des morphologues, des typologues et prévoient l'acquisition (semi)automatique de lexiques étiquetés morphologiquement, ce qui permettra l'injection de connaissances morphologiques et de ressources lexicales dans l'analyse syntaxique, elle-même optimisée ;
2) vont tirer profit de leur appartenance & pilotage du consortium TGIR – IR Corpus Ecrits, notamment axe 7 'prise en compte nouveaux modes de communication', pour être en mesure de réaliser efficacement l'analyse de la Toile « spontanée », y identifier les néologismes, et les créations lexicales ;
3) comptent optimiser le développement de grammaires hybrides sensibles au contexte par l'étude et l'implémentation de formalismes synchrones, et l'auto-apprentissage sélectif ;
4) démarrent l'étude sur la segmentation, la lemmatisation et l'analyse des textes bruités, voire mélangeant plusieurs langues (code-switching).

Conclusion :

● Avis global sur le thème :

Très positif : les résultats obtenus sont très significatifs et le projet est prometteur.

● Points forts et possibilités liées au contexte :

L'unité est une des références en France pour le traitement automatique de l'analyse (morpho)-syntaxique. Les outils, par ailleurs libres d'accès, ont une bonne notoriété : MElt, FRMG, BONSAI, SxPIPE, Lefff. Le transfert de technologie (partenariats industriels) est significatif : SxPipe → Kwaga. L'unité participe à des projets ANR SEQUOIA et Edylex ainsi qu'au Labex EFL.

Le projet visant l'analyse du contenu spontané de la toile est primordial pour la communauté des chercheurs (ainsi que pour l'industrie, pour d'autres raisons). Il devrait permettre à l'unité d'impulser un groupement de travail avec des spécialistes de disciplines diverses s'intéressant à ce médium : lexicologues, morpho(-phono)logues, pragmaticiens, énonciativistes, sociolinguistes…

● Points à améliorer et risques liés au contexte :

Il faudrait améliorer l'ergonomie des outils pour des utilisateurs ayant des compétences en informatique limitées et pour des linguistes sans compétence en TALN (pour lesquels les ressources sont également difficiles à manipuler).

En ce qui concerne les analyseurs, il faudrait mieux analyser et rendre lisible les similitudes ainsi que les différences entre les outils et/ou ressources. La complémentarité de chacun, leurs rôles et leurs impacts ne sont pas toujours très clairs : il conviendrait de préciser la place (et la priorité) de chacun dans la chaîne de traitement dit de bas niveau.

● Recommandations :

Il faudrait définir les priorités concernant les outils à développer, en considérant la disponibilité des participants, et surtout dans la perspective du démarrage de ces nouveaux projets.

Il conviendrait aussi de prévoir des programmes à plus long terme, à visée recherche (autour de la toile, par exemple).

Thème 2 : Sémantique et Discours

Effectifs : toute l'unité

• Appréciations détaillées

Il s'agit d'un thème qui monte en puissance dans l'unité, c'est une suite assez logique et conforme à un phénomène général au niveau international.

Le sous-thème de la sémantique lexicale est centré sur le développement de l'analyse de la relation prédicat argument, en continuité et en conformité avec les travaux dans ce domaine de par le monde. La priorité est de développer des ressources pour le français, avec l'objectif qu'elles soient adaptées aux besoins et utilisées. Les relais en sont les projets WOLF (wordnet libre du français) et, en projet, ASFALDA.

Ce travail se concrétise par le développement de la relation syntaxe sémantique via le French TreeBank sur laquelle l'unité ALPAGE a une bonne expertise.

Enfin, ce travail sera à l'origine du développement d'un analyseur robuste en sémantique de surface.

Le sous-thème du discours est caractérisé par le développement de ressources pour le français (lexique des connecteurs, LexConn), l'analyse du discours et de la factivité (en particulier finaliser les D-STAGs pour le discours, projet POLYMNIE), l'analyse de phénomènes temporels (concrétisé par la ressource Timebank). L'étude des phénomènes temporels est liée à celle que ces phénomènes entretiennent avec la sémantique lexicale. Enfin, sur un plan plus théorique, une action vise l'analyse des relations de discours et de leurs contraintes.

Conclusion :

● Avis global sur le thème :

Ce thème est bien organisé et se positionne de façon intéressante, prometteuse et utile pour la communauté.

● Points forts et possibilités liées au contexte :

Le projet en sémantique lexicale concerne le développement d'une base de données (« FrameNet ») du français (ASFALDA) : il s'agit de développer une sémantique fine des relations entre le prédicat et ses dépendants (arguments, ajouts). Il s'agit d'une sémantique générale, cette perspective est utile pour de nombreuses applications qui ne demandent pas une analyse très profonde des énoncés et de catégories complexes (telles que les adverbes ou les adjectifs).

Ce projet aura un impact sur l'analyse de la coopération syntaxe-sémantique.

Enfin, l'un des points majeurs est le développement de coopérations industrielles. Ceci est associé à un excellent potentiel de recherche du coté des étudiants qui peuvent largement contribuer à cet effort important de recherche.

● Points à améliorer et risques liés au contexte :

Il est nécessaire de rendre plus lisible le positionnement aux niveaux national et international sur des activités à forte concurrence qui sont susceptibles de collaborations (par exemple, FrameNet). En complément, il peut être central de contribuer à approfondir des dimensions théoriques en émergence dans ces cadres (FrameNet et l'analyse temporelle, ainsi que l'analyse du discours par les TAGs).

Une réflexion sur la notion d'analyse profonde devrait être menée: en effet, par delà les travaux sur la structure prédicat / arguments, l'unité pourrait développer des liens et des interactions avec des disciplines proches, par exemple : sémantique formelle, logique et représentation des connaissances et raisonnement ou bien encore les sciences cognitives.

- Recommandations :

Les projets sont de qualité, toutefois nous suggérons en premier lieu de définir des priorités compte tenu de l'ampleur affichée des tâches à réaliser. Corrélativement, il est nécessaire de préciser les collaborations et leurs apports concrets. Enfin, il nous parait utile pour l'avenir de l'unité de veiller à développer une recherche plus fondamentale sur le discours, où, comme il a été souligné, il reste beaucoup de travail de fond à réaliser.

Thème 3 : Linguistique quantitative et applications industrielles

Effectifs : toute l'unité

• Appréciations détaillées

Relativement à ce thème, l'unité articule son positionnement applicatif autour de deux axes : le premier est la linguistique 'expérimentale' (i.e. l'extraction de connaissances et l'évaluation d'hypothèses linguistiques à partir de l'observation de grandes quantités de données) et le second est l'ingénierie linguistique (le développement de ressources et d'outils destinés à la communauté scientifique et aux applications industrielles).

L'unité travaille en priorité sur la langue française, ce qui est à souligner et à saluer, d'autant que le laboratoire adopte une approche active de mise à disposition des ressources produites auprès de l'ensemble de la communauté scientifique.

Quoique souvent difficile et lourde à porter, cette démarche volontariste est bien réussie (ALEDA, NOMOS). Notons aussi un nombre significatif de contacts industriels et de projets menés à leur terme (SAPIENS, LIBELLEX, etc.), et la création de la start-up Verbatim Analysis. Par ces collaborations, il semble s'effectuer un véritable transfert d'expertise du laboratoire vers le monde de l'entreprise, en particulier les PME.

Conclusion :

● Avis global sur le thème :

D'une excellente qualité, les travaux ont des retombées visibles et multiples, notamment la production et la valorisation de données très conséquentes sur le français et des opportunités de financement pour les étudiants.

● Points forts et possibilités liées au contexte :

Il est à souligner un forte implication dans des projets ANR (Passage, Sequoia, Edylex, Rhapsodie, Scribo, etc.), qui se concrétise par une production très importante de ressources linguistiques ainsi qu'une excellente visibilité industrielle dans les secteurs couverts par le laboratoire. L'insertion du laboratoire dans le Labex EFL constitue une opportunité majeure d'accroître son rayonnement et son impact.

● Points à améliorer et risques liés au contexte :

Il convient d'améliorer la lisibilité et la traçabilité de l'utilisation des ressources produites. Il est aussi essentiel de faciliter leur accès aux non-informaticiens ainsi que d'évaluer l'adéquation de l'offre des ressources par rapport aux attentes de la communauté (sans pour autant aligner systématiquement l'offre sur la demande).

D'un point de vue méthodologique, il serait intéressant de mieux mettre en valeur les aspects méthodologiques qui ne sont pas spécifiques à la langue française et qui pourraient être transposés à d'autres langues, y compris dans des contextes applicatifs.

● Recommandations :

La stratégie applicative du laboratoire est cohérente et mérite d'être poursuivie. Tout en restant ouvert à de nouvelles opportunités, le laboratoire se doit d'être vigilant sur la définition de priorités pour ne pas se retrouver en situation d'éparpillement et conserver une cohérence dans la politique partenariale et l'exploitabilité scientifique des résultats.

Les perspectives de débouchés applicatifs ouvertes par le volet du projet de recherche centré sur l'analyse des textes bruités (ou spontannés, issus du web en particulier) sont particulièrement pertinentes et d'un important intérêt scientifique.

De même, il nous semble tout à fait judicieux de bien valoriser les résultats obtenus en analyse 'symbolique' en approfondissant leur intégration applicative avec les approches orientées statistiques, dans le but d'obtenir une suite d'analyseurs très performants, et fédérant ainsi les compétences présentes dans l'unité.

5 • Déroulement de la visite

Dates de la visite :

Début : 4 Décembre 2012 à 9h30

Fin : 4 Décembre 2012 à 17h30

Lieu(x) de la visite : Rue Clisson, Paris 13

Institution : Universite de Paris 7 D. Diderot

Adresse : 13, rue Clisson Paris

Locaux spécifiques visités : laboratoire

Déroulement ou programme de visite :

09h30-10h00 : **Accueil puis Réunion des membres du comité**

10h00-10h15 : **Introduction de la visite par le délégué AERES**
Présence : membres du Comité, représentants des tutelles, délégué AERES, tout ou partie de l'unité

10h15-11h45 : **Présentation du bilan et du projet de l'unité par le directeur de l'unité**
Présence : membres du Comité, représentants des tutelles, délégué AERES et/ou tout ou partie de l'unité

11h45-12h30 : **Démonstrations**
Présence : membres du Comité, représentants des tutelles, délégué AERES, toutou partie de l'unité

14h00-15h00 : **Rencontre avec les représentants du personnel (ITA, étudiants, postdocs, chercheurs)** *Présence : membres du Comité, délégué AERES, sans la direction de l'unité et sans les responsables d'équipe*

15h00-15h30 : Réunion du comité avec les représentants des tutelles
Présence : membres du Comité et délégué AERES

16h00-17h30 : **Réunion avec les directeurs (actuels, futurs) de l'unité puis**
Réunion du comité à huis clos
Présence : membres du Comité, **avec le délégué AERES et sans les tutelles**.

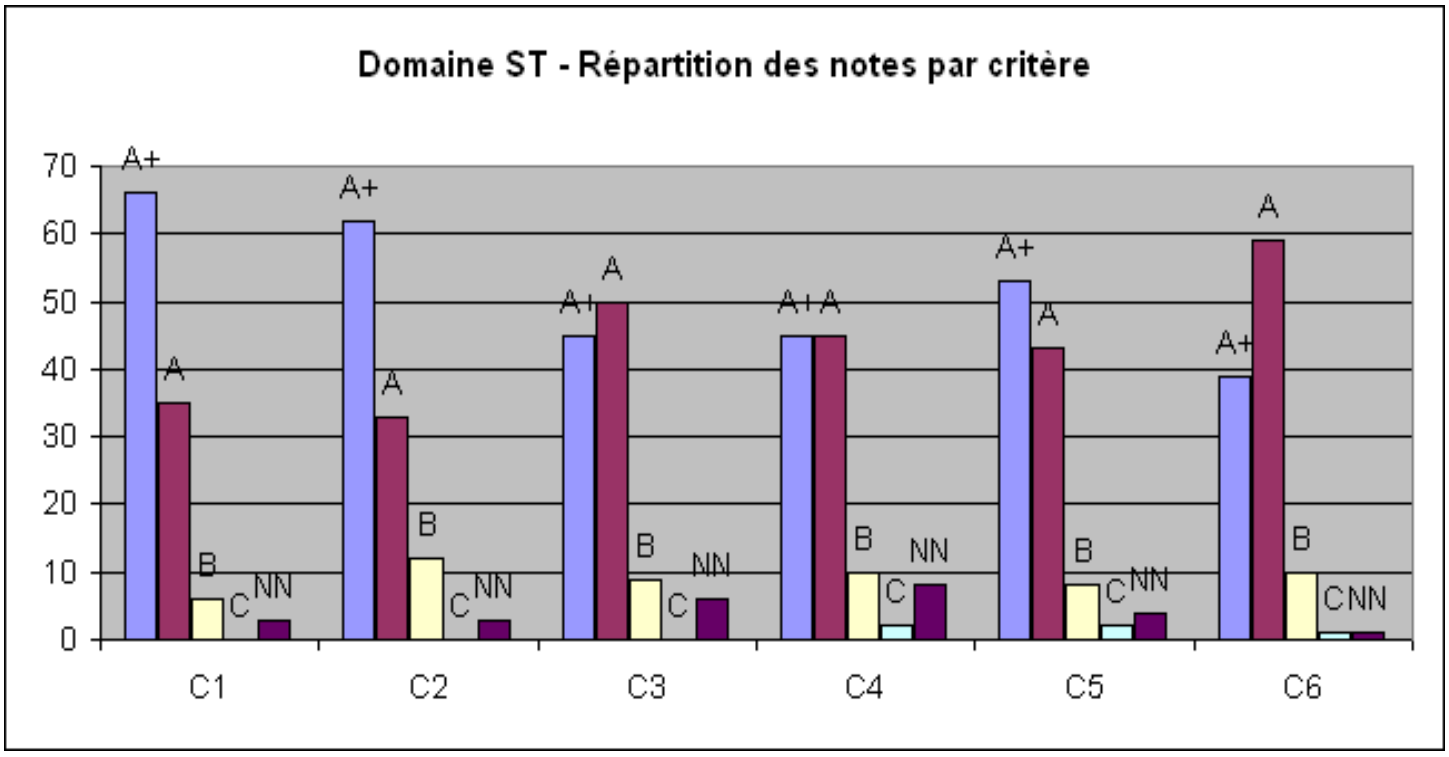
6 • Statistiques par domaine : ST au 10/06/2013

Notes

Critères	C1 Qualité scientifique et production	C2 Rayonnement et attractivité académiques	C3 Relations avec l'environnement social, économique et culturel	C4 Organisation et vie de l'entité	C5 Implication dans la formation par la recherche	C6 Stratégie et projet à cinq ans
A+	66	62	45	45	53	39
A	35	33	50	45	43	59
B	6	12	9	10	8	10
C	0	0	0	2	2	1
Non Noté	3	3	6	8	4	1

Pourcentages

Critères	C1 Qualité scientifique et production	C2 Rayonnement et attractivité académiques	C3 Relations avec l'environnement social, économique et culturel	C4 Organisation et vie de l'entité	C5 Implication dans la formation par la recherche	C6 Stratégie et projet à cinq ans
A+	60%	56%	41%	41%	48%	35%
A	32%	30%	45%	41%	39%	54%
B	5%	11%	8%	9%	7%	9%
C	0%	0%	0%	2%	2%	1%
Non Noté	3%	3%	5%	7%	4%	1%



7 • Observations générales des tutelles

Inria
INVENTEURS DU MONDE NUMÉRIQUE

université PARIS DIDEROT PARIS 7

P/VB/RL/NC/YM – 2013 - 081
Paris, le 19 avril 2013

M. Pierre Glaudes
Directeur de la section des unités de l'AERES
20 rue Vivienne
75002 PARIS

S2PUR140006360 - Analyse Linguistique Profonde À Grande Échelle - ALPAGE - 0751723R

Monsieur le Directeur,

Nous tenons, en premier lieu, à remercier les membres du comité de visite de l'AERES pour la production du rapport sur la situation de l'unité de recherche ALPAGE, rapport très élogieux, qui souligne la très grande qualité de la recherche qui y est produite, attestée par le haut niveau qualitatif et quantitatif des publications, son attractivité, l'excellente intégration des doctorants dans l'unité et sa capacité à nourrir des partenariats avec le monde industriel.

L'Université et l'INRIA réfléchiront ensemble aux moyens à mobiliser pour soutenir la dynamique engagée au sein de l'unité sur les projets de traitement automatique du langage naturel.

Je vous prie d'agréer, Monsieur le Directeur, l'expression de toute ma considération.

Vincent Berger
Président de l'Université Paris Diderot

Isabelle Ryl
Directeur du centre de recherche
Inria Paris - Rocquencourt

Tél +33 (0) 1 57 27 55 10
Fax +33 (0) 1 57 27 55 11
secretariat.president@univ-paris-diderot.fr
www.univ-paris-diderot.fr

Adresse Postale
Présidence
Grands Moulins
75205 Paris Cedex 13

SORBONNE PARIS CITÉ
RECHERCHE ENSEIGNEMENT SUPÉRIEUR

Paris, le 12/04/13

Réponse d'ALPAGE sur le rapport d'évaluation AERES

Commentaires généraux

Les membres d'ALPAGE remercient le comité d'évaluation pour son travail et son rapport. Les quelques point signalés ci-dessous relèvent de légers désaccords qui peuvent venir d'un manque de clarté dans notre rapport.

L'aspect pluri-disciplinaire de l'équipe est quelque peu passé sous silence. Il est longuement question des compétences et travaux en linguistique de l'équipe, spécialisée en TAL (Traitement Automatique des Langues), mais rien n'est dit de nos travaux en algorithmique et grammaires formelles qui sont pourtant essentiels dans la stratégie même de l'équipe et qui nous permettent de traiter des données textuelles volumineuses (passage à l'échelle, robustesse et adaptation aux domaines).

Concernant les compétences et travaux en linguistique, notre recherche en TAL repose très largement sur des connaissances linguistiques approfondies, mais il ne fait pas partie du cœur de métier d'ALPAGE de faire de la recherche en linguistique descriptive et théorique (recherche menée par exemple au sein du Laboratoire de Linguistique Formelle de Paris-Diderot). Nous participons cependant à des travaux de modélisation que ce soit en morphologie, en syntaxe (coordination) ou en discours (factualité). Voir aussi nos travaux en linguistique expérimentale (thème 3 ci-dessous).

L'ambition d'ALPAGE est de développer une chaîne de traitement complète qui couvre tous les niveaux du langage, de la morphologie à la pragmatique, et ce, dans différentes langues (le français principalement mais aussi les langues à morphologie riche) et dans différents styles (du langage châtié au langage de la Toile). Le projet même d'ALPAGE s'étale sur le long terme et nous sommes donc surpris qu'il nous soit reproché de n'avoir que des projets à court et moyen terme.

Commentaires sur le thème 1 (Morphologie et syntaxe)

Nous pensons tout d'abord que les thématiques des "langages formels" et de la "linguistique formelle" sont suffisamment disjointes pour constituer deux directions

Bureaux :
30, rue du Château des Rentiers
F – 75013 Paris

Adresse postale :
Université Paris 7 – UFRL case 7003
F – 75205 Paris Cedex 13

Tél. : +33 (0) 1 57 27 57 66
Fax : +33 (0) 1 57 27 57 81
http://alpage.inria.fr/

distinctes dans le bilan. En effet, les travaux sur les langages formels relèvent de problématiques d'informatique théorique. À l'inverse, les travaux en linguistique formelle relèvent de problématiques linguistiques, qui couvrent également des travaux en linguistique quantitative et en linguistique expérimentale, qui sont et seront effectués en grande partie dans le contexte du LabEx EFL.

De plus, la seconde "innovation majeure", qui porte sur le domaine de la morphologie lexicale et computationnelle, n'est pas limitée au "passage au contenu de la Toile non contrôlée". Elle vise également à proposer des modèles et des outils permettant la description et l'évaluation quantitative de systèmes morphologiques pour des langues typologiquement diverses, en collaboration avec des spécialistes de linguistique descriptive, typologique et formelle. À ce titre, ces travaux relèvent directement de la linguistique quantitative et constituent un programme de recherche trans-disciplinaire innovant et prometteur. Cette perspective de long terme vient en plus de celle concernant l'analyse de données bruitées issues de la Toile et de l'intégration entre analyse syntaxique et connaissances lexicales (morphologiques, syntaxiques, sémantiques ou acquises par des méthodes distributionnelles).

Commentaires sur le thème 2 (Sémantique et discours)

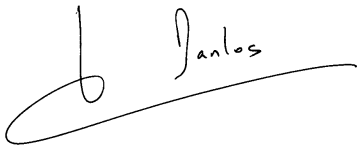
Le comité recommande de développer des collaborations internationales dans le cadre de l'utilisation du modèle FrameNet. Nous tenons à préciser que nous sommes et avons été en contact depuis le montage du projet ASFALDA avec d'autres initiatives à l'étranger (avec le projet FrameNet à Berkeley et le projet allemand SALSA).

Le comité suggère également d'améliorer les liens avec les disciplines proches : sémantique formelle, raisonnement, sciences cognitives. Nous faisons remarquer que ces liens seront tout à fait naturels lorsqu'ALPAGE aura des résultats concrets en analyse sémantico-discursive de surface, mais que nous ne pouvons pas nous permettre aujourd'hui de couvrir un aussi large spectre en sémantique.

Nous voudrions par ailleurs insister sur deux objectifs qui nous semblent être des points forts de ce thème, et qui n'ont peut-être pas été assez mis en avant :
- coupler modélisation intra-phrastique et inter-phrastique (discursif),
- développer un modèle joint d'analyse semi-supervisée syntaxique et sémantique, en collaboration avec le LIF dans le cadre du projet ASFALDA.

Commentaires sur le thème 3 (Applications)

Dans le rapport d'ALPAGE, le thème 3 recouvre d'une part un axe linguistique empirique et expérimentale et d'autre part l'axe ingénierie linguistique (y compris les applications industrielles). Or le rapport d'évaluation passe presque sous silence les activités en linguistique empirique et expérimentale. Pourtant, cette thématique a fait l'objet de deux thèses soutenues et de plusieurs mémoires de Master, ainsi que de nombreuses publications dont plusieurs articles de revues. Enfin, cette thématique occupe à ce jour une dimension fondamentale en linguistique théorique et descriptive comme en témoigne la création du Labex EFL dont ALPAGE est un acteur important.



Laurence Danlos
Responsable d'ALPAGE
Professeur Université Paris Diderot