préGDR TAL
Axe de réflexion : Compréhension

Maxime Amblard[1] et Marianna Apidianaki[2]

[1]Campus Scientifique, BP 239
54506 Vandoeuvre-lès-Nancy
maxime.amblard@univ-lorraine.fr

[2]rue John von Neumann, 91403 Orsay
LIMSI, CNRS, Université Paris-Saclay
marianna.apidianaki@limsi.fr

12 juin 2018

Ce document présente la synthèse de la réflexion sur le thème Compréhension du préGDR TAL. Les éléments présentés recensent des activités de recherche menées en France autour de ce sujet, et plus particulièrement en Sémantique, pendant les 10 dernières années. Ils resument également les avis exprimés par des spécialistes du domaine qui ont été convoqués afin de fournir leur opinion par rapport à l'état actuel des recherches en Sémantique, et de faire des propositions concrètes pour l'avenir de la recherche en Sémantique en France.

1 Thèmes et périmètre d'étude

La compréhension du langage naturel est un vaste domaine de recherche dont l'objectif est de fournir aux modèles computationnels les capacités nécessaires pour interpréter le sens de textes, afin d'améliorer l'expérience des utilisateurs lors de leur interaction avec les machines. Les applications en sont nombreuses et vont de la classification des textes jusqu'aux systèmes question-réponse, l'extraction d'information, la traduction, et les systèmes de dialogue. La compréhension consiste particulièrement à l'identification de concepts, de catégories sémantiques et de leurs relations, au repérage des événements et des actions exprimées dans des textes, ainsi que des acteurs impliqués, à l'identification des intentions et des sentiments exprimés par les utilisateurs. Elle implique aussi la combinaison d'éléments linguistiques et extra-linguistiques – liés à la situation de communication – afin de situer le contenu des textes sur un axe temporel, permettre un ancrage linguistique, et établir le lien entre les concepts exprimés

et le monde. Cette tâche complexe comprend donc un grand nombre de sous-tâches précises pouvant être centrées sur un niveau de traitement particulier (par exemple, le traitement sémantique ou syntaxique), ou sur l'interaction entre des niveaux d'analyse différents (par exemple, l'interface syntaxe-sémantique ou sémantique-pragmatique).

La compréhension dans une acception générale recouvre des problématiques variées qui rélèvent de domaines de connaissance différents (de la modélisation et la compréhension du fonctionnement du cerveau (au sens physiologique) au positionnement philosophique du rapport aux savoirs, en passant bien évidemment par la langue naturelle). Ces questions de recherche s'appuient sur, et se servent souvent, des théories et représentations développées dans les domaines de la Linguistique et du Traitement Automatique de la Langue (TAL). Le pont par lequel il nous est naturel de passer est donc celui de la sémantique, prise comme l'étude de la signification des mots et de leur combinaison afin de créer, et d'interpéter, des énoncés cohérents. Evidemment, la compréhension ne se limite pas à l'analyse sémantique des énoncés linguistiques, mais concerne également l'établissement de liens avec la cognition aussi bien qu'avec le contexte textuel et situationnel plus large, et le monde.

Nous identifions deux cadres théoriques dans lesquels s'inscrivent les recherches en sémantique computationnelle.

— La **sémantique formelle**, qui s'appuie sur une tradition symbolique et qui est centrée autour des propriétés inférentielles de la langue [7, 8, 2]. Elle s'appuie sur la philosophie de la langue et exploite des mécanismes inférentiels issus de la logique.

— La **sémantique distributionnelle**, qui, quant à elle, est statistique et dirigée par les données, et principalement focalisée sur des aspects sémantiques liés au contenu descriptif [4, 6, 3]. La sémantique distributionnelle peut s'intéresser aussi bien au sens des éléments lexicaux isolés, qu'à leur combinaison au niveau des énoncés ou à celui des textes pour former un discours cohérent.

Bien que les recherches dans ces deux cadres ont souvent été poursuivies de manière indépendante, ceux-ci sont complémentaires et des tentatives de combiner les forces qui émanent de chaque cadre commencent à apparaître [1]. Des collaborations sont également mises en place au niveau administratif et organisationnel, où l'on retrouve des conférences comme *SEM, [1] conjointement organisées par les SIGs (Special Interest Groups) d'ACL sur la Sémantique (SIGSEM) et le Lexique (SIGLEX). De manière analogue, des conférences traditionnellement organisées sur la sémantique formelle comme LACL [2] ou IWCS [3]

1. Des informations sur les conférences *SEM (Joint Conference on Lexical and Computational Semantics) des deux dernières années peuvent être trouvées ici : `https://sites.google.com/site/starsem2017/`, `https://sites.google.com/view/starsem2018`

2. Logical Aspects of Computational Linguistics : `http://lacl.gforge.inria.fr/lacl-2016/index-presentation.html`

3. International Conference on Computational Semantics : `http://iwcs2015.github.io`, `https://www.lirmm.fr/iwcs2017/`

accueillent aujourd'hui des contributions sur la sémantique distributionnelle [5].

Les moyens et ressources nécessaires pour parvenir à produire des représentations pour chacun de ces paradigmes sont riches, et varient en fonction de l'architecture des systèmes de compréhension envisagés et des tâches visées par ces systèmes. Ils comprennent :

— des **ressources** sémantiques, telles que lexiques, ontologies, taxonomies ou corpus annnotés, décrivant les concepts et leurs relations

— des **outils** permettant l'analyse des textes, comme des analyseurs syntaxiques, des systèmes de désambiguïsation lexicale, des étiqueteurs en rôles sémantiques, des outils d'inférence textuelle et d'analyse des sentiments

— de grands volumes de **données** dans la langue concernée, brutes (dans le cas de l'apprentissage non supervisé) ou enrichies avec des informations linguistiques (dans le cas de l'apprentissage supervisé)

Dans la suite du document, nous synthétisons les avancées qui ont eu lieu en France pendant les dix dernières années en termes de systèmes, de ressources et d'outils visant la compréhension du français, et comment ceux-ci se sont concrétisés dans le cadre de projets européens et français menés sur le territoire. Basés sur des opinions fournies par des experts dans le domaine en France et à l'étranger, nous fournissons également un aperçu général des problèmes et lacunes présents dans ce domaine en France, et nous faisons des propositions concrètes pour faire avancer la recherche dans ce domaine en France dans les années à venir.

2 État des lieux

La **sémantique formelle** s'appuie, depuis de nombreuses années, sur le travail d'informaticiens théoriciens et logiciens. Il y a un vrai savoir faire français sur la question, mais avec une communauté de très petite taille au regard de ce qui existe dans d'autres pays européens, par exemple en Allemagne. Cependant cette branche est en perte de vitesse ces dernières années, la thématique n'intéressant plus explicitement les informaticiens théoriciens et logiciens. Dans le même temps, les théories de la sémantique formelle ont plutôt été reprises pour être développées vers le discours et la cognition. De fait, la question du développement d'outils à large couverture n'a pas été un sujet majeur pour la communauté. Cette absence de ressources a conduit à un isolement relatif de ces recherches au regard de la dynamique apportée par les méthodes distributionnelles et les méthodes par apprentissage profond, ainsi qu'à un manque de visibilité d'un point de vue applicatif et industriel.

La **sémantique distributionnelle** a, au contraire, bénéficié d'une vraie dynamique dans la période récente. De très nombreux résultats ont été apportés par les méthodes d'apprentissage (supervisé, non-supervisé ou profond). L'un

des problèmes pour ce type de paradigme est de parvenir à avoir un modèle explicatif des résultats produits. Si pour certaines tâches cela n'est pas primordial, pour la question de la compréhension et celle de la sémantique cela reste crucial. C'est donc un enjeu particulier pour ces méthodes que de parvenir à une solution pertinente. De manière étonnante, cette diversité de résultats ne se cristalise pas autour d'une solution homogène. L'une des difficultés qui semblent émerger est l'absence de ressources en commun, bien que la question du multilinguisme soit relativement bien couverte.

La sémantique n'a été consolidée ni par la branche formelle ni par la branche distributionnelle. La communauté dispose d'un nombre important de collègues apportant des résultats pertinents, mais peu visibles. Il n'existe pas d'école française de la sémantique à l'international, et finalement les liens au niveau national sont très peu structurés. De nombreux collègues s'accordent sur la nécessité de parvenir à une hybridation entre les deux types d'approche. Nous voyons d'un côté une bonne couverture sur les aspects théoriques et formels, de l'autre de vraies compétences et résultats obtenus en utilisant des méthodes d'apprentissage automatique, mais très peu dans l'interval.

L'une des conséquences intéressantes est que la communauté s'est penchée sur les questions du parsing sémantique et du parsing du discours, qui tendent plus vers la notion de compréhension. Ces tâches permettent de réintroduire dans la problématique d'autres communautés scientifiques, d'autres perspectives, ou encore d'autres cas d'application. Cet intérêt est important, mais présente le même risque que celui de la sémantique : d'un côté voir émerger le développement de théories formelles non pourvues d'outils, de l'autre le développement de méthodes d'apprentissage s'appuyant difficilement sur les structures linguistiques et n'apportant pas toujours une interprétation. La question des ressources pour la sémantique et le discours se retrouve de fait au cœur de la problématique des deux aspects. Cela tend également à pointer l'importance de la sémantique lexicale et des liens qu'elle entretien avec le reste de la communauté.

Nous fournissons en Annexe A une liste des laboratoires français qui s'activent dans le domaine de la sémantique formelle et distributionnelle, et de l'apprentissage pour la sémantique. L'Annexe B recense les projets français et européens (impliquant des équipes françaises) qui ont été menés au cours des dix dernières années, et qui avaient un lien avec la sémantique.

La première remarque est la présence d'une quinzaine d'équipe de recherche sur la problématique. Les forces sont donc largement présentes en France. On note également une véritable diversité de disciplines, de l'informatique traditionnelle à la linguistique formelle, voire la psycholinguistique et la cognition. Cependant, peu de groupes ne semblent concernés par mobiliser la sémantique et les processus de compréhension dans le cadre des sciences cognitives expérimentales. L'ensemble de ces groupes couvrent naturellement l'ensemble du territoire.

Du point de vue des projets de recherche, l'ANR a financé depuis 10 ans une cinquantaine de projets sur le TAL, dont une trentaine sur la connaissance dans une acceptation large. Seulement une dizaine de projets entre dans le cadre de

notre définition de la compréhension. Ce nombre est relativement important et il montre une véritable dynamique sur la problématique dans l'interval, les projets s'étalant sur la période 2007-2017. On note également que les deux paradigmes sont présents. De la même manière, les projets se sont intéressés tout autant à la constitution de ressources qu'au développement d'outils et de méthodes pour les analyser.

La problématique est également présente sur des projets de grande envergure que sont les ERC de la communauté européenne, et les IUF. L'Annexe B récapitule l'ensemble de ces projets.

Du côté des ERC, on peut noter le projet de François Recanati en 2008 sur le contexte, le contenu et la compositionnalité, ainsi que le projet de Nicholas Asher sur les stratégies de conversation en 2010. Par ailleurs, deux projets en psycholinguistique et sémantique formelle ont été acceptés en 2012, avec comme investigateurs principaux Philippe Schlenker et Emmanuel Chemla. Pour les IUF, on peut noter les projets de Benoît Crabbé en 2014, Laurence Danlos en 2013 et Francis Corblin en 2010. De manière générale, si la thématique a été bien pourvue au début des années 2010, elle semble en perte de représentants en cette fin de décénie, ce qui va dans le sens de nos précédentes analyses.

3 Perspectives

Cette section recense les avis d'experts en Sémantique Computationnelle et Traitement Automatique des Langues en France et à l'étranger, récoltés via un questionnaire établi en ligne. Les noms des personnes ayant répondu au questionnaire sont donnés en Annexe C.

De manière générale, la communauté constate un manque en ressources nécessaires pour l'entraînement et l'évaluation de modèles sémantiques, l'absence de communication et de collaboration entre les équipes, ainsi que le peu de partage des outils développés.

Ressources La communauté constate un manque de ressources sémantiques à grande échelle pour le français, nécessaires pour entraîner et évaluer des modèles statistiques. Ce manque limite le développement de systèmes de traitement sémantique et de compréhension pour cette langue. Les ressources nécessaires comprennent des corpus annotés à grande échelle avec des informations sémantiques, ainsi que des jeux de données encodant des connaissances comme les sens des mots et leurs relations, la paraphrase, l'inférence textuelle et les rôles sémantiques. Les ressources sémantiques (lexiques, dictionnaires, réseaux sémantiques) de bonne qualité actuellement disponibles pour le français sont soit payantes, ce qui limite leur utilisation, soit faibles en couverture, ce qui pose des contraintes pour le traitement de texte libre. Les ressources automatiquement créées sont souvent de qualité moyenne, ce qui rend leur utilisation dans des applications difficile. En outre, la disponibilité de corpus annotés en français est très faible, même si des efforts de transfert inter-langue de connaissances linguistiques de différentes sortes depuis des corpus anglais annotés ont vu le jour ces dernières

années, ce qui offre une alternative à la constitution coûteuse de corpus annotés dans d'autres langues.

Il y a également un besoin de partage des outils et modèles développés au sein des différentes équipes, actuellement pas disponibles pour être réutilisés par d'autres groupes. Le développement de ressources sémantiques à grande échelle aidera le développement de systèmes de traitement sémantique (en France et à l'étranger). Le partage des outils facilitera la comparaison des résultats obtenus, ce qui fera avancer la recherche en sémantique du français, et favorisera la collaboration entre équipes.

Méthodes et modèles La sémantique formelle (approches sémantiques basées sur la logique) est bien développée et représentée en France, avec une communauté bien identifiable. D'autres approches (sémantique distributionnelle, dirigée par les données, basée sur les connaissances) sont répresentées dans différentes équipes mais de manière sporadique et diffuse. Aussi, des équipes "leader" en apprentissage profond pour la sémantique sont difficiles à identifier.

Il y a actuellement besoin d'intégrer les travaux de recherche antérieurs en sémantique avec l'apprentissage profond. Combiner les approches symboliques et statistiques avec l'apprentissage profond constitue un réel défis pour la communauté. La sémantique computationnelle intégrant des méthodes formelles existe mais peine dans le passage à l'échelle, préférant se focaliser sur le discours plutôt que sur le parsing sémantique. En outre, bien que les deux pourraient bénéficier des importantes avancées dans les domaines de la psycholinguistique et des sciences cognitives, le peu de collaboration entre les communautés rend rare l'exploitation des données, des ressources et des résultats dans un cadre interdisciplinaire.

Les approches distributionnelles ou neuronales pourront également profiter de l'intégration de connaissances externes pour générer des représentations sémantiques de meilleure qualité. En outre, vu le manque de ressources à grande échelle pour le français, il y a besoin de modèles généraux capables d'apprendre à généraliser à partir de peu de données. Nous devons alors se diriger vers des modèles plus robustes, généraux et reproductibles.

Pour favoriser la compréhension, la sémantique doit également être vue sous différents angles, allant au-delà du mot et des expressions multi-mots, vers la génération de représentations sémantiques de textes. L'interaction entre différents niveaux d'analyse linguistique (comme l'interface entre syntaxe et sémantique) doit aussi être explorée, et les travaux multilingues (comme les approches de transfert de connaissances sémantiques) doivent être maintenus et encouragés. Finalement, il est important de développer des modèles sensibles au contexte pouvant soutenir des agents conversationnels adaptatifs.

Collaborations Afin de profiter des avancées en sémantique en France dans le nouveau contexte mondial imposé par le succès des modèles d'analyse profonde, il faudra favoriser le travail interdisciplinaire entre des experts en apprentissage automatique et les experts en linguistique et en Traitement Automatique des

Langues orienté vers des applications. Actuellement, on constate une bonne couverture d'aspects théoriques et d'aspects d'apprentissage automatique en France, mais pas beaucoup de travaux se situant entre les deux.

On constate également une déconnexion entre les approches computationnelles et expérimentales en France, avec très peu de cas de collaboration entre les deux. Les recherches en apprentissage et acquisition des langues, menées dans les domaines de la psycholinguistique et des sciences cognitives, peuvent fournir des données hautement utiles pour le design et la mise en place de systèmes d'analyse statistique et d'apprentissage profond.

Finalement, il y a besoin d'établir le lien entre la sémantique et la communication multi-modale afin de permettre aux représentations développées en sémantique d'être exploitées dans des agents conversationnels afin d'améliorer la communication avec les humains.

Applications On constate souvent un cloisonnement des recherches en sémantique, et l'absence de lien direct avec les applications. Néanmoins, les modèles et représentations développés en sémantique pourraient faciliter et améliorer le traitement dans des applications précises (par ex. dans les systèmes de dialogue, les systèmes questions-réponses, la traduction automatique, l'extraction d'information). Pour que les modèles sémantiques puissent être exploités dans des applications réelles, les aspects computationnels qui permettraient aux modèles de passer à l'échelle et de traiter de grands volumes de texte libre doivent être considérés.

4 Programmatique

4.1 Communauté

Il est important de structurer la communauté par des collaborations dans deux directions. D'une part, il est essentiel de constituer une communauté française de la sémantique, en encourageant la communication entre des équipes travaillant en sémantique formelle et distributionnelle, mais aussi entre des experts en sémantique et des équipes abordant des questions cognitives plus larges ou liées à l'apprentissage. D'autre part, il est important de parvenir à ce que la communauté de sémantique s'articule avec d'autres communautés du TAL (extraction d'information, web sémantique, parole, dialogue, traduction), afin de proposer des modèles et des représentations sémantiques à grande échelle et pouvant être intégrées dans des applications réelles. Cette dynamique peut s'accompagner par l'organisation de volumes thématiques dans des journaux nationaux et internationaux.

La création d'un groupe de travail au sein du GDR et l'établissement d'une journée thématique annuelle, permettraient d'augmenter la visibilité des recherches menées dans les différentes équipes, et encourageraient les discussions entre les différentes communautés. L'organisation de séminaires et de groupes de travail pourrait également contribuer à établir des collaborations et à mettre

en place des projets communs.

Plus précisément, il serait intéressant d'organiser un séminaire national virtuel, en passant par les systèmes de visio-conférence et Renater. Le programme serait monté par le GDR et les créneaux établis très en amont afin de garantir une large participation.

4.2 Ressources

Etant donné le besoin pressant en ressources annotées pour le français pour faire avancer les recherches qui impliquent notamment des techniques d'apprentissage supervisé et profond, il est primordial de mettre en place une charte pour le partage des ressources disponibles, et d'encourager le développement de nouvelles ressources à grande échelle pour le français.

Les ressources peuvent encoder des connaissances sémantiques en forme de lexiques, mais peuvent aussi consister en des jeux de données annotées pour le français. En outre, il est important de partager des outils, modèles et métriques d'évaluation. Ce point est particulièrement important car il permettra la réproducibilité des résultats obtenus auparavant. Il permettra également de mettre en place des protocoles d'évaluation communs pour différentes tâches, hautement utiles pour la comparaison des systèmes et l'établissement de l'état de l'art. La charte de partage de données et d'outils, et les protocoles d'évaluation pourraient s'inspirer des protocoles équivalents disponibles dans d'autres langues, et surtout de ceux mis en place au sein de campagnes d'évaluation internationales (par ex. la campagne SemEval, qui regroupe de nombreuses tâches d'évaluation pour des systèmes effectuant différents types de traitements sémantiques). [4]

Afin de faciliter le partage entre différentes équipes, une plateforme commune pourrait être établie pour accueillir des données, des modèles, des outils, et des scripts d'évaluation pour différentes tâches.

4.3 Tâches d'évaluation

La mise en place de tâches d'évaluation pour le français permettrait également la communication entre les équipes, augmenterait la visibilité des recherches menées au sein de celles-ci, et créerait une dynamique pour la recherche en sémantique en France. Afin de développer des tâches d'évaluation, il faudra développer des jeux de données pour le français couvrant différents phénomènes, tâches, problématiques et besoins. De telles données sont disponibles pour plusieurs tâches en anglais et quelques autres langues (au sein des campagnes SemEval), mais le français n'y est pour l'instant pas présent sauf dans des rares cas de tâches multilingues. [5] Les tâches SemEval pourraient fournir une base pour la mise en place de tâches focalisées sur le français, qui pourraient néanmoins

4. Le descriptif des tâches proposées dans le cadre des campagnes SemEval des deux dernières années est accessible ici : `http://alt.qcri.org/semeval2017/`,`http://alt.qcri.org/semeval2018/`

5. Un tel cas est la tâche Cross-lingual Semantic Parsing with UCCA à SemEval 2019 (`http://alt.qcri.org/semeval2019/index.php?id=tasks`).

s'en différencier pour couvrir des questions pertinentes spécifiques pour cette langue.

La participation à ces tâches est possible à titre individuel ou au niveau des équipes. Afin de motiver la communauté à y participer, il serait important d'encourager la formation de groupes de travail sur des thématiques précises entre différents laboratoires et équipes. Il faut noter que ces tâches d'évaluation pourraient également servir comme une base pour des travaux dirigés ou pour la mise en place de projets pour les étudiants dans un cadre universitaire, comme cela commence à se faire dans les universités anglophones pour des tâches SemEval.

L'objectif principal de la mise en place de ce type de tâches, serait toutefois de fournir une base commune pour l'évaluation et la comparaison des systèmes développés pour le français, et d'augmenter la visibilité des résultats au niveau international.

4.4 Formation et enseignement

L'une des conséquences du manque de structuration nationale de la thématique se répercute sur la question de l'enseignement et de la formation. Les étudiants sont rarement formés au TAL spécifiquement, et d'autant moins aux questions de la compréhension. On peut identifier quelques propositions qui restent trop singulières.

Le GDR doit pouvoir aider à agencer un programme s'adressant aux étudiants en thèse, par exemple sous la forme d'école d'été spécialisée (la thématique pourrait tourner entre les différents thèmes du GDR d'année en année). Cette école serait également l'occasion de revenir sur la difficulté de faire travailler ensemble des spécialistes de thématiques différentes pour, par exemple, trouver des accords terminologiques.

Le transfert de la recherche vers la formation est primordial pour attirer les étudiants vers ces thématiques et former les futurs acteurs du domaine au meilleur niveau. Intégrer les nouvelles approches et technologies utilisées dans le domaine au niveau international au sein des formations dispensées dans les universités pourrait être un objectif du GDR.

Nous proposons deux idées pour alimenter la réflexion. Comme introduit dans la section précédente, l'organisation de tâches s'appuyant sur la constitution de ressources pourrait également servir à définir des exercices appliqués à destination des Masters. Enfin, il nous parait important d'encourager la mobilité des étudiants entre les laboratoires français, et ce dès le Master puis en Thèse.

4.5 Médiation scientifique

Une confusion est largement répandue auprès du grand public sur la capacité réelle des systèmes à répondre à une tâche de compréhension automatique. Souvent réduite à la stricte reconnaissance de la parole, il est important d'identifier des systèmes, des exemples, des cas d'étude mettant spécifiquement en avant les difficultés de la compréhension.

Il est important de diffuser auprès des médias scientifiques, voire des médias grand public. Le GDR pourrait expliciter auprès des collègues les attendus de la médiation scientifique, par exemple par la rédaction d'articles abordables. On devrait également encourager la participation des équipes de recherche à des événements nationaux comme la Fête de la Science, [6] qui favorisent les échanges entre la communauté scientifique et le grand public.

6. https://www.fetedelascience.fr

Références

[1] Gemma Boleda and Aurélie Herbelot. Formal Distributional Semantics : Introduction to the Special Issue. Computational Linguistics, 42(4) :619–635, December 2016.

[2] Robin Cooper and Christian Retoré. An outline of type-theoretical approaches to lexical semantics. Journal of Language Modelling, 5(2) :165–178, 2017.

[3] Katrin Erk, Diana McCarthy, and Nicholas Gaylord. Measuring Word Meaning in Context. Computational Linguistics, 39(3) :511–554, 2013.

[4] Zellig Harris. Distributional structure. Word, 10(23) :146–162, 1954.

[5] Reinhard Muskens and Mehrnoosh Sadrzadeh. Context update for lambdas and vectors. In International Conference on Logical Aspects of Computational Linguistics, pages 247–254. Springer, 2016.

[6] Peter D. Turney and Patrick Pantel. From Frequency to Meaning : Vector Space Models of Semantics. Journal of Artificial Intelligence Research, 37(1) :141–188, January 2010.

[7] Piek Vossen. Handbook of formal languages. Journal of Logic, Language and Information, 14(4) :457–487, Oct 2005.

[8] Yoad Winter. Elements of formal semantics : An introduction to the mathematical theory of meaning in natural language. Edinburgh University Press, 2016.

Annexe A

Forces

Nous avons constitué une liste des laboratoires et équipes françaises actives dans le domaine de la Compréhension, et où des recherches plus particulièrement en Sémantique sont menées. Cette liste est non exhaustive, et est susceptible à être étendue.

- Laboratoire de Science Cognitive et Psycholinguistique [Le langage et son acquisition Modéliser le développement cognitif] (UMR 8554)
- Laboratoire de Linguistique Formelle (LLF) [Grammaire de l'énoncé et du discours] (UMR 7110)
- Analyse et Traitement Informatique de la Langue Française (ATILF) [Discours](UMMR 7118)
- Laboratoire lorrain de recherche en informatique et ses applications (Loria) [Sémagramme, Synalp, Orpailleur, Smart, Multi-Speech] (UMR7503)
- Institut de Recherche en Informatique de Toulouse [MEthodes et ingénierie des Langues, des Ontologies et du DIscours (Melodi)] (UMR 5505)
- Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM) [Texte] (UMR 5506)
- CEA Tech [List]
- Laboratoire d'Informatique de Grenoble (LIG) [GETALP (Groupe d'Étude en Traduction Automatique/Traitement Automatique des Langues de la Parole)] (UMR 5217)
- Langues, Textes, Traitements informatiques et Cognition (LATTICE)(UMR 8094)
- Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur (LIMSI) [Traitement du langage parlé, écrit et gestuel] (UPR3251)
- Institut de recherche en informatique et systèmes aléatoires (IRISA) [LINKMEDIA] (UMR 6074)
- Groupe de recherche en informatique, image, automatique et instrumentation de Caen (GREYC) [HULTECH] (UMR UMR6072)
- Laboratoire Parole et Langage (LPL) (UMR 6057)
- Laboratoire d'Informatique Fondamentale de Marseille (LIF)[TALEP](UMR 7279)
- Almanach, INRIA

Annexe B

Projets ANR

Depuis 10 ans, 50 projets sur le TAL ont été financés par l'ANR. Parmi ces projets, on en identifie une trentaine qui portent sur la sémantique et le discours.

- LISE : Linguistique, normes, traitement automatique des langues et Sécurité : du « data et sense-mining » aux langues contrôlées (2007)
- DEMOCRAT : DEscription et MOdélisation des Chaînes de Référence : outils pour l'Annotation de corpus (en diachronie et en langues comparées) et le Traitement automatique (2015)
- Hybride : Hybridation de la fouille de données et du traitement automatique des langues(2011)
- TALAD : Analyse et traitement automatique de discours (2017)
- CABeRneT : Compréhension Automatique de Textes Biomédicaux pour la Recherche Translationnelle. (2013)
- MULTISEM : Modèles Avancés pour le Traitement Sémantique Multilingue (2016)
- ANNODIS : Annotation discursive : corpus de référence pour le français et outils d'aide à l'annotation et à l'exploitation(2007)
- ASFALDA : Analyse Sémantique en Frames : Annotation, Lexique, Discours et Automatisation (2012)
- Sensunique : Optimisation d'un logiciel pour la rédaction de textes techniques de qualité : application-pilote au domaine de la santé.(2010)
- TermITH : TERMinologie et Indexation de Textes en sciences Humaines (2012)
- Polymnie : Analyse et synthèse dans les grammaires catégorielles abstraites : du lexique au discours (2012)
- WEB-NLG : Génération de texte pour le web sémantique (2014)
- ASRAEL : Acquisition de Schémas pour la Reconnaissance et l'Annotation d'Événements Liés (ASRAEL) (2016)
- PARSEME-FR : Analyse syntaxique et expressions polylexicales pour le français (2015)
- DATCHA : Extraction de connaissances à partir de vastes corpus de conversations "chat" client-opérateurs (2015)
- GoAsQ : Modélisation et résolution de requêtes ontologiques sur des don-

nées médicales semi-structurées (2016)

— ContentCheck : Techniques de gestion de contenus pour la vérification des faits : modèles, algorithmes et outils (2016)

— PIITHIE : Plagiat et Impact de l'Information Textuelle recHerchée dans un contexte InterlinguE (2006) reprendre les projets ANR/ERC/IUF

— Semantiques langue

— Polycat : La polyvalence catégorielle : un paramètre linguistique universel ? Approches grammaticales, sémantiques et cognitives (2006)

— Vagueness : Cognitive Origins of Vagueness (2007)

— LOCI : Locativité et Interaction en Logique, Linguistique et Informatique 52010)

— PRELUDE : Vers une pragmatique théorique basée sur la théorie des continuations et sur la ludique (2006)

— GRASP : Apprentissage automatique par les graphes pour la prédiction de structures linguistiques (2016)

— Comprendre : Compréhension du discours et des événements (2008)

— CoFee : Feedback Conversationel : Analyse et Modélisation multi-dimensionnelles (2012)

— GENIUS : Genericity : Interpretation and Uses (2008)

— MINDPROGEST : Rôle de l'attribution des états mentaux dans la construction du sens : Marqueurs de référence, prosodie et geste (2011)

— Readers : Évaluation et développement de systèmes d'analyse et de compréhension de textes (2012)

— Comprendre : Compréhension du discours et des événements (2008)

Projets ERC

ERC			
nom	pomotion	statut	thématiques
Francis René, Julien Bach	Cons. 2016	INRIA	Seqoia Robust algorithms for learning from modern data
Philippe Schlenker	Ad. 2012	CNRS	Frontsem New Frontiers of Formal Semantics
Emmanuel Chemla	St. 2012	CNRS	SEMEXP Psycho-semantics : new data for formal semantics models, stronger frameworks for experimental studies
Nicholas Asher	Ad. 2010	CNRS	STAC Strategic Conversation
Francois Xavier Alario	St. 2010	CNRS	LIPS Lexical information processes and their spatio-temporal dynamics
François Recanati	Adv. 2008	CNRS	CCC Context, Content, and Compositionality
Fanny Meunier-Hoen	St 2007	CNRS	SPIN Natural speech comprehension : Comprehension of speech in noise

Projets IUF

IUF			
nom	promotion	statut	thématiques
Benoît CRABBÉ	Junior 2014	U. Paris Diderot - Paris 7	TAL
Laurence DANLOS	Senior 2013	U. Paris Diderot - Paris 7	TAL
Francis CORBLIN	Senior 2010	U. Paris-Sorbonne - Paris 4	Sémantique
Sylviane CARDEY-GREENFIELD	Senior 2008	U. de Franche-Comté	Linguistique, TAL
Laurent BESACIER	Junior 2012	U. Joseph Fourier	Parole, traduction

Annexe C

Sondage

Nous avons mis en place un sondage en ligne pour l'Axe Compréhension du Pré-GDR TAL, [7] et nous avons invité des spécialistes du domaine en France et à l'étranger à répondre à un ensemble de questions. Un court descriptif de l'objectif du sondage a été fourni, et des questions ouvertes ou à choix multiples ont été proposées.

Descriptif : The goal of this survey is to assess the current trends in Semantics research in France, and to collect the opinions of researchers and other actors on the present and future of the field.

Questions :

1. I am not interested or don't have time to fill in this survey
   but wish you good luck !
2. (a) Name (optional)
   What Semantics-related topics are addressed in your research ?
   (b) How have these topics evolved in the past ten years ?
   — Same research topics, same methodology
   — Same research topics, different methodology
   — New research topics
   (c) What is your opinion about Semantics research in France ? Which topics or areas are well-covered, which should be further developed and how do these compare to international trends ?
   (d) In your opinion, what are the key challenges in the field in France and abroad ?
   (e) What means and actions are required in order to promote these developments, support Semantics research in France and increase visibility ?

Participants : Nous remercions les 18 personnes qui ont répondu, dont : Nicholas Asher, Marco Baroni, Marine Carpuat, Vincent Claveau, Mathieu Constant, Benoît Crabbé, Emmanuel Dupoux, Gregory Grefenstette, Thierry Poibeau, Preslav Nakov, Christian Retoré, Didier Schwab.

Trois personnes n'ont pas répondu mais nous ont souhaité bonne chance.

7. Le formulaire du questionnaire est accessible à cette adresse : `https://docs.google.com/forms/d/1yCs0CUUiTgI0SZaRcyX1dfgHaqAm35gny9cUrVNEhDM/edit`