Demande de création d'un groupement de recherche

Motivations de la demande

Au cours des dernières décennies, le domaine de la linguistique informatique a connu des avancées majeures grâce notamment à l'utilisation et à l'adaptation, aux données orales et aux données écrites, de méthodes symboliques (théorie des langages formels, grammaires d'unification, théorie de la preuve), de différents modèles d'apprentissage automatique (approches génératives et discriminantes) et des méthodes neuronales (réseaux convolutionnels, approches de type encodeurs-décodeurs, etc.).

Si ces avancées ont jusque-là majoritairement servi à développer des applications (moteur de traduction automatique, systèmes question/réponse, dialogue humain/machine, détection d'opinion, transcription automatique de la parole, etc.), elles peuvent également être mises à profit pour faciliter l'analyse linguistique et plus spécifiquement pour créer des modèles falsifiables[1], pour vérifier leurs prédictions et pour découvrir des généralisations. Ainsi par exemple, les techniques de reconnaissance de la parole peuvent être exploitées pour faciliter le travail de transcription des linguistes de terrain ou pour les assister dans l'inventaire des phonèmes d'une langue non documentée ; les techniques de traduction automatique et d'alignement pour faciliter la création de lexiques bilingues nécessaires à la documentation d'une langue ; et les algorithmes d'analyse et de génération pour valider des hypothèses syntaxiques et lexicales.

Plus généralement, la linguistique informatique met à la disposition des linguistes un large éventail de techniques et de ressources qui ouvrent des perspectives nouvelles pour l'analyse linguistique, que ce soit pour collecter et annoter des données ou pour extraire ou vérifier des généralisations linguistiques. L'objectif de ce GDR est (i) d'explorer ce potentiel en tirant parti d'un réseau scientifique favorisant les interactions entre linguistes, linguistes de terrain et linguistes informaticiens et (ii) de favoriser l'émergence de méthodes nouvelles qui bénéficient à la fois aux linguistes (automatisation des processus d'analyse et de validation), aux linguistes de terrain (facilitation des processus de collecte et d'analyse des données) et aux linguistes informaticiens (développement de nouvelles techniques

[1] Un modèle linguistique est falsifiable s'il peut être contredit par les données. Par exemple, une grammaire informatique génère un langage donné. Si elle génère une phrase agrammaticale, la théorie encodée dans cette grammaire est falsifiée. On dit alors que la grammaire sur-génère. Inversement, quand la grammaire échoue à générer une phrase bien formée du langage qu'elle décrit, on dit que la grammaire sous-génère.

nécessitées par l'analyse linguistique, essor des méthodes non ou faiblement supervisées pour l'analyse des langues peu dotées, peu écrites ou non documentées).

Ces trois sous-communautés linguistiques sont bien représentées en France.

Grâce en particulier, à un soutien fort et continu du CNRS et à l'existence de plusieurs UMRs spécifiquement dédiées à l'étude et à la documentation des langues du monde (LACITO, LLACAN, SEDYL, DDL), la linguistique de terrain est très active. La linguistique formelle et la linguistique informatique sont également bien représentées avec notamment le CLLE, l'IJN, LLF, le LLING, le LLL, le LPL, MoDyCo et le LPP pour la linguistique formelle ; et l'ATILF, le LATTICE, le LIS, le LIG, le LIMSI, l'IRIT, le LORIA et STIH pour la linguistique informatique.

Comme en témoignent notamment le nombre et la qualité des manifestations (co)organisées, le domaine bénéficie d'une bonne visibilité nationale et internationale.

À ***l'interface entre linguistique informatique et analyse des langues peu dotées ou non documentées,*** différentes initiatives et projets ont récemment émergé qui attestent d'un intérêt marqué pour cette thématique. À l'international, l'atelier bi-annuel SLTU ("Spoken Language Technologies for Under-resourced languages") se tiendra pour la sixième fois en 2018. L'atelier « Computational Methods for Endangered Languages » alterne entre la conférence de l'ACL (Association for Computational Linguistics) et la conférence ICLDC (International Conference on Language Documentation and Conservation) dans le but de promouvoir les interactions entre linguistes informaticiens et linguistes de terrain. En France, un Colloque intitulé « Computational Methods for Endangered Language Documentation and Description » a récemment été organisé à l'Ecole Normale Supérieure par Thierry Poibeau (LATTICE, CNRS) en collaboration avec des collègues allemands et la fédération TUL (Typologie et Universaux Linguistiques) organise régulièrement des séminaires à l'intersection entre linguistique formelle et linguistique de terrain (e.g., le Workshop sur les outils d'élicitation pour la linguistique descriptive et la typologie organisé en novembre 2017).

En ***linguistique descriptive***, une école d'été annuelle (Fieldling, anciennement IWSLF et STL), organisée par plusieurs labos de typologie-description (LACITO, LLACAN, SEDYL) en partenariat avec Paris 3 et l'INALCO, rassemble chaque année depuis 2009 plusieurs dizaines de Masters et de doctorants français et étrangers autour des techniques de collecte et d'analyse des données propres à la linguistique de terrain et cette formation intègre régulièrement un module consacré au traitement informatique des données.

En ***linguistique informatique***, la conférence annuelle TALN (Traitement Automatique des Langues) s'accompagne depuis une dizaine d'année d'ateliers sur les langues dites peu dotées (sous diverses appellations comme DILITAL en 2017) ou les langues d'Afrique (TALAF http://talaf.imag.fr/). Un workshop sur l'interaction de la linguistique formelle et la

sémantique distributionnelle a été organisé à Toulouse en 2015 par M. Abrusan, N. Asher et T. Van de Cruys (https://www.irit.fr/semws2015/) et une école d'été du CNRS intitulée "Nouvelles Technologies pour l'Exploration de Corpus de Parole" destinée aux linguistes souhaitant acquérir des connaissances dans le traitement automatique de grands corpus oraux et co-organisée par le LIMSI, le LPP et le LIUM avec le soutien du Labex EFL aura lieu du 9 au 12 juillet 2018. Les projets BULB (Breaking the Undocumented Language Barrier), RESTAURE (Computational Resources and Processing for Regional Languages) et COPAL (Corpus Parallèles pour l'Alsacien) travaillent au développement de méthodes informatiques pour le traitement de langues peu ou non dotées. Enfin, le Zero Resource Speech Challenge organisé par E. Dupoux encourage le développement d'approches permettant l'apprentissage non supervisé d'unités linguistiques à partir de données orales.

La ***linguistique formelle*** bénéficie elle aussi d'une dynamique forte. Ainsi le Colloque biannuel de syntaxe et sémantique à Paris (CSSP) est devenu un pôle important de discussion en linguistique formelle en Europe.[2] Divers ateliers ont été récemment organisés sur la sémantique formelle (colloque "NISM 2016. New Ideas in Semantics and Modelling" organisé par C. Beyssade, A. Mari et D. Nicolas en 2016 ; Workshop `Modality, actions and events' à l'ENS, Paris en Mai 2013 ; Workshop `How do we know what happens' à l'ENS, Paris en Mai 2014). La "12th International Conference on Computational Semantics" (IWCS) a été organisée à Montpellier en septembre 2017 et l'école d'été "European Summer School on Logic, Language and Information" (ESSLLI) à Toulouse en Juillet 2017.

Objectifs généraux et opérations thématiques

Le GDR vise à dynamiser les liens entre linguistique, linguistique de terrain et linguistique informatique. La linguistique informatique et le traitement automatique des langues (TAL) ont connu ces dernières décennies des avancées scientifiques et technologiques majeures qui ont permis le développement d'applications toujours plus performantes (traduction automatique, systèmes de question/réponse, reconnaissance automatique de la parole, résumé automatique, etc.). En revanche, la place de la linguistique dans cette évolution est limitée. Les linguistes ignorent souvent l'existence d'outils et de méthodes qui leur permettraient de faciliter, d'accélérer ou de conforter leurs analyses. Inversement, les résultats des systèmes de TAL sont souvent analysés sur la base de mesures qui ne prennent que très indirectement en compte la qualité linguistique des résultats produits. Enfin les techniques de TAL reposent en grande partie sur des données annotées et de grande taille. Pour l'analyse de langues peu dotées, il importe de développer de nouvelles méthodes, moins gourmandes en données comme en annotations. Ainsi l'interaction entre linguistique informatique et linguistique de terrain devrait non seulement faciliter le travail des linguistes de terrain (avec les outils les plus récents, quatre heures de parole transcrite suffisent à entraîner un modèle de reconnaissance de la parole tout à fait correct) mais aussi

[2] Plusieurs membres du GDR proposé ont été impliqués dans l'organisation de ce colloque, ou au niveau de président du comité scientifique (Abeillé: 2013; Crysmann: 2017) ou bien du comité d'organisation (Bonami: 2009; Crysmann: 2013).

favoriser des avancées dans le domaine de la linguistique informatique (en stimulant les recherches dans le domaine de l'apprentissage faiblement supervisé). .

L'objectif du GDR est d'explorer et de promouvoir les interactions entre ces différentes dimensions de la linguistique. Il s'agit d'explorer dans quelle mesure les outils, les méthodes et les ressources créés par les recherches menées dans les domaines du TAL peuvent être exploités pour comprendre la structure de la langue ; ce que la linguistique peut apporter au développement des systèmes de TAL ; et inversement, dans quelle mesure les outils et techniques du TAL peuvent contribuer à faciliter les recherches des linguistes et linguistes de terrain, notamment en leur permettant des analyses d'une plus grande masse de données (annotation manuelles complétées par des annotations automatisées).

Le GDR sera structuré autour de cinq grands axes:

1. Extraction de généralisations linguistiques par des méthodes informatiques
2. Linguistique et évaluation des systèmes de traitement automatique des langues
3. Outils de collecte et d'analyse pour les linguistes
4. Données et défis partagés pour une science ouverte
5. Linguistique informatique pour les langues peu dotées ou non documentées

Axe 1 : Extraction de généralisations linguistiques par des méthodes informatiques

La linguistique a pour objet l'étude de la langue. Elle vise notamment à extraire des généralisations et à identifier des invariants à l'intérieur d'une langue ou entre langues d'une même famille, ainsi que des universaux (syntaxiques, sémantiques, morphologiques et phonétiques). Lorsque des données sont disponibles, les méthodes d'apprentissage automatique et d'apprentissage profond exploitées dans le domaine du TAL peuvent permettre d'apprendre ces invariants ou d'extraire des généralisations de façon automatique ou tout au moins, de proposer une première analyse qui pourra guider le linguiste dans son étude. Ainsi par exemple, la phylogénétique computationnelle permet de créer, à partir de données informatisées sur les cognats, des arbres phylogénétiques[3]. De même les ressources multilingues annotées (e.g., Universal dependencies) créées par la communauté du TAL peuvent être utilisées, en combinaison avec des méthodes d'apprentissage, pour assister les typologues dans le travail qui consiste à classifier les propriétés des systèmes linguistiques et à établir des régularités de variation linguistique en fonction de critères appris à partir des données.

[3] Longobardi G., Buch A., Ceolin A., Ecay A., Guardiano C., Irimia M., Michelioudakis D., Radkevich N. and Jaeger G. (2016). Correlated Evolution Or Not? Phylogenetic Linguistics With Syntactic, Cognacy, And Phonetic Data. In S.G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér and T. Verhoef (eds.) The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11). Available online: http://evolang.org/neworleans/papers/162.html

Un point important concerne l'interprétation des modèles produits par les réseaux de neurones. Ces modèles sont difficiles à appréhender, mais le fait qu'ils fournissent à l'heure actuelle les meilleurs résultats sur un certain nombre de tâches (par exemple en parsing ou en traduction automatique) rend nécessaire une exploration approfondie de leur contenu, même si celui-ci ne se laisse pas lire aussi facilement qu'un modèle symbolique ou même un modèle produit par des méthodes d'apprentissage classique. On peut imaginer tirer des informations précieuses d'une meilleure connaissance de ces modèles, avec potentiellement des conséquences sur le plan linguistique[4]. De même, la notion de modèles multilingues a récemment montré son potentiel et on peut y voir deux intérêts complémentaires : d'une part identifier des points de convergence entre les langues (sur la base des proximités et regroupements opérés automatiquement par le système d'analyse) et d'autre part ces approches multilingues sont très efficaces pour traiter des langues peu dotées (par exemple pour mettre au point un analyseur syntaxique efficace pour une langue sans données annotées). Enfin, les modèles issus de l'approche distributionnelle sur la sémantique lexicale ouvrent également des perspectives intéressantes. Elles peuvent permettre par exemple d'étudier l'influence du contexte sur le sens des expressions lexicales ou encore d'explorer, à travers les langues, la distinction entre expressions fonctionnelles de classe fermée dites ``universelles'' et expressions de classe ouverte.

Ce premier axe vise d'une part, à ***promouvoir l'utilisation des méthodes de TAL dans la découverte d'invariants et de généralisations linguistiques*** et d'autre part, à ***stimuler le développement de méthodes informatiques permettant l'extraction de généralisations à partir de données rares***. L'apport d'autres domaines de l'informatique à la linguistique pourra également être pris en compte comme par exemple, celui de la théorie des types, conçue pour la mathématique constructive et utilisée dans les langages de programmation pour la sémantique des langues naturelles (voir les travaux de Robin Cooper en Suède, Zhaohui Luo en Angleterre ou Nicholas Asher en France); celui de la méthode des continuations issue de l'informatique théorique pour la formulation d'une sémantique dynamique prenant en compte le contexte (voir les travaux de Philippe de Groote, Sylvain Pogodalla à Nancy) ou encore celui de la théorie des jeux utilisée en vérification de programmes pour la modélisation de la structure conversationnelle. On pourrait même ajouter un quatrième thème de l'informatique pertinente à la linguistique: la robotique a fait beaucoup de progrès et est sur le point d'intégrer de façon importante des données conversationnelles avec des données visuelles, une tâche qui intéressera tout linguiste intéressé dans la communication située ou communication gestuelle.

Axe 2 : Linguistique et évaluation des systèmes de traitement automatique des langues

L'analyse des systèmes de TAL est le plus souvent soit quantitative (une ou plusieurs métriques sont utilisées pour évaluer la qualité des résultats du système sur un ensemble de données de test) soit manuelle (des juges humains notent manuellement les résultats produits selon un protocole préétabli). Elle est souvent également guidée par la tâche plutôt

[4] Montavon, Samek and Müller 2017. Methods for interpreting and understanding deep neural networks. Digital Signal Processing 73. 1–15.

que par la qualité linguistique des résultats produits. Par exemple, la sortie d'un système de traduction automatique est comparée à un ensemble de traductions de référence et si elle partage un nombre suffisant de segments avec cet ensemble, elle obtient un bon score même si, par ailleurs, elle contient des fautes de syntaxe ou d'orthographe. La conception de systèmes réellement utilisables exige cependant de pouvoir garantir la qualité linguistique des résultats obtenus.

Un premier objectif du GDR sera ***d'explorer dans quelle mesure des critères linguistiques peuvent être utilisés pour évaluer, de façon automatique, la qualité linguistique des résultats (traductions, résumés, etc.) produits par les systèmes de TAL***.

Un second objectif sera de c***onstruire, à partir des recherches conduites en linguistique formelle, des batteries de tests linguistiques qui permettent d'évaluer les systèmes de TAL.*** Il existe déjà des premiers efforts visant à analyser quels types de généralisations linguistiques sont extraites par les méthodes d'apprentissage automatique contemporaines. Par exemple, (Linzen et al. 2016) a examiné dans quelles conditions un modèle de langue neuronal parvient à modéliser les contraintes d'accord. De même, (Isabelle et al. 2017)[5] propose un jeu de tests permettant d'évaluer la capacité des systèmes de traduction automatique à traiter de différences morphologiques, lexicales ou syntaxiques entre langue source et langue cible. Des progrès substantiels dans cette direction ne seront possibles qu'à partir d'une systématisation des découvertes de linguistique fondamentale dans une forme appropriée pour le traitement informatique (jeux de tests, tâches partagées, défis scientifiques, etc.).

Axe 3 : Outils de collecte et d'analyse pour les linguistes

L'objectif de cet axe est (i) de concevoir des outils permettant de faciliter le travail des linguistes et des linguistes de terrain et (ii) de promouvoir l'utilisation d'outils, de ressources et de techniques existants.

Par exemple, pour documenter les langues en voie de disparition, il est urgent de pouvoir collecter et traduire les données orales de ces langues. Afin de faciliter, accélérer et améliorer ce processus, différents outils informatiques ont récemment été proposés par des linguistes informaticiens telle l'application mobile AIKUMA[6] développé par Steven Bird qui permet d'enregistrer la parole spontanée ainsi que la traduction et la répétition, à un rythme plus lent, de ces enregistrements[7] ainsi que l'extension LIG-AIKUMA développée par le LIG à Grenoble qui inclut un mode "Correction" permettant au linguiste de corriger du texte (erreurs orthographiques, syntaxiques, de prononciation, etc.) et un mode «Élicitation » permettant d'éliciter de la parole auprès du locuteur au moyen de textes, d'images ou encore

[5] P. Isabelle, C. Cherry and G. Foster. A Challenge Set Approach to Evaluating Machine Translation - https://arxiv.org/pdf/1704.07431.pdf

[6] Bird S., Gawne L., Gelbart K. and McAlister I. Collecting Bilingual Audio in Remote Indigenous Communities. COLING, 2014 - aclweb.org

[7] La "re-dite" vise à faciliter la transcription *a posteriori* des données enregistrées.

de vidéos. Exploitant la puissance et la légèreté des téléphones portables, ces logiciels libres permettent, d'une part, de collecter des données orales de bonne qualité, et d'autre part, d'associer ces données à des textes numériques (traduction, transcription) directement utilisables par des processus informatiques en aval, comme par exemple, l'alignement texte/parole, mais également, dans le cas où la taille des données est suffisante, la détection automatique ou semi-automatique des éléments constitutifs d'une langue (phonèmes, morphèmes, mots formes, unités lexicales, expressions phraséologiques, structures grammaticales, etc.).

Pour donner un autre exemple, il devient de plus en plus commun en linguistique théorique comme en linguistique expérimentale d'utiliser les plongements des mots et des phrases. Ces plongements sont des vecteurs numériques dérivés automatiquement des corpus de textes qui comprennent diverses informations sémantiques sur les mots. Ils permettent par exemple d'étudier les corrélations entre variantes dérivationnelles et variation sémantique, de créer des classes sémantiques ou encore de vérifier des hypothèses sur les relations lexicales. Des plongements existent pour plus de 40 langues (https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1989) mais ils ne sont pas facilement utilisables par les linguistes. Une plateforme telle que rusvectores.org qui permet à chacun de consulter les propriétés des plongements des mots et de les manipuler a de nombreuses applications possibles pour le linguiste : elle lui permet par exemple de rechercher les voisins d'un mot dans l'espace sémantique, de calculer les distances entre les mots, de résoudre des analogies sémantiques ou encore de visualiser les mots dans un espace distributionnel. Une plateforme qui intègre des fonctionnalités similaires et des plongements pour différentes langues permettrait de faciliter l'utilisation des plongements par les linguistes.

Un troisième exemple concerne l'homogénéisation des plateformes d'outils d'annotation, et l'aide à l'annotation ou l'enrichissement des annotations pour lesquelles de grands progrès restent possibles. Même dans une plateforme dédiée telle qu'ELAN, la transcription requiert toujours un temps absolument considérable (plusieurs centaines d'heures de travail pour la transcription d'un corpus de quelques heures dans une langue peu décrite est une moyenne). Le recours à des outils tiers comme FLEx ou (The Field Linguist's) Toolbox, ou encore Praat pour l'annotation phonétique/phonologique est indispensable pour améliorer le rendement du travail des annotateurs, mais leur intégration à la plateforme de documentation prédominante, ELAN, est imparfaite (ces outils ne supportent pas les formats multimédia aujourd'hui généralisés dans le travail de terrain, et posent de multiples problèmes d'incompatibilité selon les plateformes et les versions). On peut parler à cet égard d'une vraie difficulté à procéder à une annotation aidée. La mise au point d'outils réellement intégrés, couvrant le plus grand nombre possible des aspects de l'annotation aidée (transcription phonémique des données orales[8], annotation lexicale, morphologique et

[8] Oliver Adams, Trevor Cohn, Graham Neubig, Alexis Michaud. Phonemic transcription of low-resource tonal languages. Wong, Sze-Meng Jojo ; Haffari, Gholamreza. *Australasian Language Technology Association Workshop 2017*, Dec 2017, Brisbane, Australia. ISSN: 1834-7037, pp.53-60, 2017, Australasian Language Technology Association Workshop 2017: Proceedings of the workshop. 〈http://alta2017.alta.asn.au/alta2017-draft-proceedings.pdf〉

phonologique) faciliterait le travail des linguistes et ouvrirait d'autant mieux la voie à des traitements automatisés ultérieurs.

Plus généralement, de nombreux outils et techniques développés dans le domaine du TAL peuvent contribuer à accélérer la collecte des données mais également à produire des données de meilleure qualité pour l'analyse linguistique. L'objectif de ce troisième axe est de ***faire connaître ces outils et ces techniques et de promouvoir leur utilisation par les linguistes et les linguistes de terrain*** (par exemple au cours de séminaires et/ou écoles d'été pluri-disciplinaires). A l'inverse, ***les échanges entre les linguistes confrontés aux exigences de la description de terrain et les informaticiens pourront contribuer à améliorer des outils disponibles ou même à créer de nouveaux outils à partir des besoins exprimés***.

Axe 4 : Données et défis partagés pour une science ouverte

Afin de promouvoir les interactions entre membres du GDR, cet axe du projet visera à collecter les données de travail des membres du GDR dans un espace partagé ; à favoriser le partage et la réutilisabilité des données linguistiques (batteries de tests linguistiques, données orales et écrites collectées par les linguistes de terrain) ; à faciliter et à standardiser les processus d'annotation des données ; à définir et à organiser des défis scientifiques communs ; ainsi qu'à définir, à partir de ces données et autour des thématiques abordées au sein du GDR, des thématiques de travail communes comme par exemple, le traitement multi niveaux (oral, lexique, morphologie, syntaxe) d'une langue non documentée. A travers ces actions, l'objectif est de ***promouvoir, dans la communauté linguistique, une culture de publication où le partage des données devient un élément intégral de la communication scientifique, et où la réutilisation des données est reconnue par citation*** obligatoire de la publication accompagnante les données. Cette culture de partage des données a déjà bénéficié à plusieures disciplines comme par exemple la bioinformatique et l'intelligence artificielle. En France, le projet BULB a récemment mis à la disposition de la communauté un corpus de données orales collectées par des linguistes de terrain sur le Mboshi (Bantu C25) afin de permettre la reproductibilité des expériences et des évaluations faites à partir de ces données[9]. Un objectif important du GDR sera de favoriser l'émergence de telles initiatives au sein de la communauté linguistique.

Axe 5 : Linguistique informatique pour les langues peu dotées ou non documentées

Les langues pour lesquelles peu de données existent, constituent un défi majeur pour le TAL en ce qu'elles exigent la conception de nouvelles méthodes d'apprentissage (apprentissage non- ou faiblement supervisé) pour lesquelles des connaissances linguistiques a priori sont

[9] https://arxiv.org/abs/1710.03501 P. Godard, G. Adda, M. Adda-Decker, J. Benjumea, L. Besacier, J. Cooper-Leavitt, G-N. Kouarata, L. Lamel, H. Maynard, M. Mueller, A. Rialland, S. Stueker, F. Yvon, M. Zanon-Boito. A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments - à paraître - LREC2018.

souvent requises. Cet axe du GDR vise à ***développer des méthodes informatiques utiles et utilisables pour le traitement de langues peu dotées***. Il s'agira par exemple, de développer des méthodes d'apprentissage faiblement, semi ou non supervisées pour des données de taille réduite provenant de langues peu dotées, peu écrites ou non documentées ; de concevoir, implémenter et tester des méthodes d'expansion automatique des données permettant d'appliquer des méthodes d'apprentissage automatique à des données de petite taille ; d'utiliser et d'adapter les méthodes symboliques (grammaires computationnelles, automates à états finis) pour l'analyse et la génération de langues peu dotées (afin par exemple de valider une grammaire et de tester sur- et sous-génération)[10] ; ou encore, d'exploiter des méthodes d'apprentissage automatique pour créer des moteurs de gloses permettant de minimiser les prétraitements.

Cet axe sera développé en lien avec l'axe « extraction de généralisations linguistiques par des méthodes informatiques » : il est fréquent que l'on ait des données et des ressources pour certaines langues et beaucoup moins de données (par exemple juste des textes non annotés) pour une langue proche ou apparentée (langue de la même famille linguistique, parfois simplement langues en contact). La recherche de généralisations et de traits partagés peut permettre d'induire des connaissances relativement poussées sur la langue peu dotée dans ce type de configuration, y compris avec des techniques demandant généralement beaucoup de données (comme les réseaux de neurones). Ce type de processus doit évidemment être contrôlé sur le plan linguistique pour ne pas généraliser indûment tout et n'importe quoi, mais il semble malgré tout précieux pour les langues peu dotées.

Modalités de fonctionnement et financement

Fonctionnement

Le GDR dans son ensemble sera coordonné par Claire Gardent et Denis Paperno, assistés par un comité scientifique qui se réunira une fois par trimestre. Chaque axe sera porté et animé par un responsable d'opération. Des activités communes ou associant plusieurs opérations seront régulièrement organisées. Un site Web implanté au LORIA permettra d'assurer la visibilité des activités du GDR et la diffusion des informations.

1. SÉMINAIRE TRIMESTRIEL DU GDR.
Un séminaire trimestriel sera organisé par le comité scientifique avec pour objectif de rendre visibles les opérations conduites dans le GDR et d'inviter des personnalités scientifiques externes (françaises ou étrangères) dont les travaux sont pertinents pour le GDR.

2. ATELIERS

[10] Par exemple, le traitement computationnelle de la morphophonologie non-concaténative, de la réduplication et de la résomption qui sont piloté dans la grammaire implémentée du haoussa (Crysmann, 2015a, 2015b, 2016) peuvent servir comme patron réutilisable pour des langues diverses.

Chaque responsable de groupe organisera des ateliers propres à assurer le développement et la visibilité de son opération. Chaque membre du GDR dans son ensemble sera tenu informé de toutes les réunions tenues par les opérations et invité à s'y rendre s'il le souhaite.

3. JOURNÉES d'ÉTUDE et COLLOQUES

Le GDR dans son ensemble et chacune des opérations seront invités à tenir des journées d'étude ou colloques ouverts, permettant à une communauté plus large que celle du GDR (notamment, la nouvelle société / conférence SCIL https://blogs.umass.edu/scil/scil-2018/, qui vise à renforcer les interactions entre linguistique et linguistique computationnelle) de présenter et de débattre de questions spécifiques. Il proposera aussi sa collaboration pour organiser des ateliers ou sessions parallèles dans les conférences où la linguistique joue un rôle (conférences de l'ACL, de l'ATALA, ICLDC, CSSP, etc.).

4. MOBILITÉ DES ÉTUDIANTS

Le GDR cherchera à faciliter la mobilité des étudiants en thèse et en Master dans le cadre soit de stages, soit de visites plus brèves. Il cherchera également à promouvoir les co-directions de thèse sur des thèmes à l'interface entre linguistique informatique et linguistique de terrain, linguistique formelle et linguistique de terrain ou encore linguistique informatique et linguistique formelle.

5. ÉCOLE D'ÉTÉ

Le GDR prévoit de tenir une école d'été « Linguistique, Linguistique de terrain et Linguistique Computationnelle ». Cette école pourrait durer une dizaine de jours. Elle associera un enseignement théorique en linguistique informatique, linguistique formelle et linguistique de terrain avec une initiation à l'implémentation de modèles à partir de données et d'outils collectés ou construits par le GDR.

6. PUBLICATIONS

Chacune des opérations du GDR cherchera à contribuer à des publications
- en recherche fondamentale: numéros de revues ou volumes
- dans le domaine de la formation des étudiants et chercheurs

Liste des projets de recherche et travaux des membres du GDR

- GDRI OASIS (2018-2021) -- *Ontology As Structured by the Interface with Semantics*. Porteurs : B. Copley, I. Roy. 15000 euros per year. SFL (Paris 8), LLF (Paris 7), LLING (Nantes), Queen Mary (UK), UPF (Spain), ZAS (Germany), UIT (Norway), MIT (US), USC (US).
- GDRI SEEPiCLa (2016-2020) -- *Structure, Emergence and Evolution of Pidgin and Creole Languages*. Scientific Coord. : E. Schang. 14000 Euros/an. Universities of Haiti, Mauritius, Lexington (USA), Westminster (UK), Amsterdam, Paris-Diderot, Paris-8, Orléans, ZAS.
- Action COST European Network for *Combining Language Learning with Crowdsourcing Techniques* (enetCollect), 2016-2021, porteuse pour la France : K. Fort (Sorbonne Université). 36 pays participants. 185.000 euros la première année.

- Projet ERC SPEECHREPORTING (2018-2023) -- *Discourse Reporting in African Storytelling.* Porteuse : Tatiana Nikitina. 1,5 million euros.
- Projet ERC STAC (2011-2017) -- *Strategic Conversation*. Advanced researcher grant no. 269427, Porteur : Nicholas Asher, 1,93 million euros
- Projet Marie-Curie TAMEAL (2009-2014) -- *The interrelation of Tense, Aspect and Modality with Evidentiality in Australian Aboriginal languages*. Grant Agreement PIRSES-GA-2008-230818-TAMEAL. Porteur : Patrick Caudal, 82800 €
- Projet Marie Curie Career Integration Grant CONTENT (2013-2017) -- *Content across domains: From words to discours*' Porteuse : M. Abrusan, 100 000 euros, IRIT, UPS,
- National Science Foundation (2014-18) -- *Language Induction meets Language Documentation: Leveraging bilingual aligned audio for learning and preserving languages*. Porteurss: David Chiang and Steven Bird. US$470k, Notre Dame University and University of California Berkeley
- Projet DFG SFB 732-B1, (2014-2018) -- *The Form and Interpretation of Derived Nominals dans le SFB 732 Incremental Specification in Context*. Chercheur: Gianina Iordăchioaia (PIs: Artemis Alexiadou: 2014-2016; Silke Fischer: 2016-2018). 270 000 euros. Institute of Linguistics and Institute of Natural Language Processing, University of Stuttgart.
- Projet Procope (PHC franco-allemand), (2017-2018) -- *One-to-Many Relations in Morphology, Syntax and Semantics*. Porteur: Berthold Crysmann, 6.000 €, LLF. Partenaire allemand (DAAD) : Manfred Sailer, U Francfort.
- Projet RoGraV (2016-2020) -- *The Grammar of the Verbal Domain in Romanian and Beyond*. Porteurs : Carmen Dobrovie-Sorin et Ion Giurgea, 92200 euros, LIA franco-roumain..
- Projet Institut Universitaire de France (IUF), 2014–2019 -- *Lexicologie multidimensionnelle : théorie, construction et exploitation des réseaux lexicaux*. Porteur : Alain Polguère. 75 000 €. ATILF CNRS.
- Projet Institut Universitaire de France (IUF junior), 2014–2019 -- *Analyse syntaxique et sémantique avec des représentations distribuées*. Porteur : Benoit Crabbé, 75000€. LLF CNRS.
- Projet fédération TUL (2015-2018) -- *Denumerals across languages*. Porteur : Bernard Fradin. 7 000. I. Bril, D. Creissels, A. Nakajima, A. Söres, F. Rose, A. Lahaussois, A. Roulois, Y. Treis, A. Vittrant
- Projet ERA-NET Atlantis (2016-2019) -- *Grounded language learning*. LATTICE. Porteur : Vrije Universiteit Brussel (VUB), project coordinator, The Austrian Research Institute for Artificial Intelligence (OFAI), The Institut de Biologia Evolutiva (IBE) at the University Pompeu Fabra in Barcelona, LATTICE, SONY CSL (paris / Japan)
- Projet ITEA2 ModelWriter (2014-2017) -- *Synchronizing Texts and Models.* Porteur : Ferhat Erata. 4 183 000 euros. Airbus, LORIA Nancy, Obeo, Unit, Ford Otosan, Hisbim, KocSystem, Mantis, Unit.
- Projet ITEA3 Papud (2017-2020) -- *Profiling and Analysis Platform Using Deep Learning.* Porteur : C. Crisan (BULL). 10 927 000 Euros. Detaysoft, Bull / ATOS, Hi-Iberia, GeneralSW, ISEP, Beia, Koc Sistem, Ericsson, KU Leuven University, Press' Innov, PERTIMM, Softeam, Mines Telecom, LORIA / University of Lorraine, Lille 1 University, Abalia, Universidad Politécnica de Madrid, Kuveyt Turk A.Ş., Türk Telekom, Assist, Performetric, Universidade do Minho.
- Projet ANR *Orfeo* (2013-2017) -- Porteuse : J.-M. Debaisieux. ATILF, CLLE-ERSS, ICAR, LATTICE, LIF, LLF, LORIA, MODICO.
- Projet ANR Democrat (2015-2019) -- *Analyse automatique des chaînes de coréférence* (français médiéval and français contemporain). Porteur : Frédéric Landragin. 385 736 euros. ICAR, LATTICE, LILPA.
- Projet ANR Profiterole (2018-2021?) -- *Analyse syntaxique de l'ancien français et diachronie* (10e-17e). Porteuse : Sophie Prévot. 371 519 Euros. ICAR, LATTICE, LLF.
- Projet ANR ALFFA (2013-2017) -- *African Languages in the Field – Fundamentals and Automation* – Porteur : L. Besacier. 400.000 Euros. LIG, LIA, DDL, Voxygen SA.
- Projet ANR-DFG BULB (2014-2018) -- *Breaking the Unwritten Language Barrier* - Porteur : Gilles Adda. 725.000 Euros. LIMSI, LIG, LPP, LLACAN, ZAS (Ge), KIT (Ge).
- Projet ANR WebNLG (2014-2018) -- *Natural Language Generation for the Semantic Web*. Porteuse : Claire Gardent. 251 935 Euros. LORIA/Nancy, Stanford Research International USA, KRDB Bolzano, Italie.Projet ANR SegCor (2015-2019) - *Segmentation of oral corpora*. PI : V.Traverso, T.Schmidt. ICAR, LLL, IDS Mannheim
- Projet ANR PARSEME-FR (2015-2020) - *Multiword expressions and syntactic parsing in French*. Porteur : M. Constant. ATILF, LI, LIS, LIFO, LLF.
- Projet ANR DATCHA (ANR-15-CE23-0003) (2015-2019) - *Knowledge extraction from large corpora of human-human conversation data from web chat services* - Aix-Marseille Univ, Institut de Recherche en Informatique de Toulouse, Orange Labs
- Projet ANR Démonext (2018-2021) -- *Dérivation Morphologique en Extension*. Porteuse : Fiammetta Namer (ATILF). 600 000 euros. Autres participants : UMR STL, CLLE-ERSS, LLF.
- Projet ANR Parsiti (2016-2021) -- *Parsing the Impossible, Translating the Improbable*. Porteur : Djamé Seddah. 530 000 Euros. Inria, LIMSI, LIPN.
- Projet ANR SoSweet (2015-2020) -- *A Sociolinguistics of Twitter: Social Networks and Linguistic Variations*, Porteur :Jean-Philippe Magué, resp. pour Inria Paris: D. Seddah. 627 000 Euros. ENS Lyon, Inria Dante, Inria Paris, Univ. Grenoble.

- Projet ANR TermITH (2012-2016) -- *Terminologie et Indexation de Textes en Sciences Humaines et Sociales*. Porteuse : Evelyne Jacquey. 710 719 euros. ATILF, INIST, INRIA Saclay, LIDILEM, LINA, LORIA.
- Projet ANR HimalCo (2013-2016) -- *Corpus parallèles en langues himalayennes* Porteur : Guillaume Jacques (CRLAO). 198 000 €. CRLAO, LACITO, HTL, Collection Pangloss.
- Projet ANR LangAge (2018-2022) -- *Differences in language learnability across ages.* co-Porteuses : A. Cristia and S. Peperkamp
- Projet ANR MechELex (2014-2018) -- *Mechanisms of early lexical acquisitio*n. Porteuse : A. Cristia €252,000l
- Projet ANR MathSegPhon (2017-2019) -- *The mathematics of segmental phonotactics.* Porteur : G. Magri (€110,000)
- Projet ANR TriLogMean (2014-2018) -- *Trivalent Logics and Natural Language Meaning*. Porteur : Benjamin Spector (co-Porteur : Paul Egré), 212000 €
- Projet ANR-NSF (2017-2021) -- *Neuro-Computational Models of Language*. Porteur : J. Hale, C. Pallier. Partenariat U. Cornell, U. Michigan, INSERM-CEA, LLF et Inria Paris.
- Projet ANR Croissant (2018-2021) -- *Les parlers du Croissant : une approche multidisciplinaire du contact oc-oïl*. Porteur: Nicolas Quint (250.000€).Projet ANR SYMILA (2013-2017) -- Syntactic Microvariation in the Romance languages of France. Porteurs : Patrick Sauzet, Anne Dagnac, Dominique Sportiche (http://symila.univ-tlse2.fr/)
- Projet PSL Lakme (2016-2018) -- *Computational methods for morphologically-rich languages* (français médiéval, hébreu et langues finno-ougriennes -- saami, komi). LATTICE.
- Projet CPER CoReA2D (2017-2019) -- *Corpus et Ressources : Annotation, Apprentissage et Désambiguïsation*. Porteuse : Evelyne Jacquey. 24 000 euros. ATILF, INIST
- Projets Stativité / StaTyC / ETA-Tyc (ATILF/ Région Lorraine / UL) (2015-2017) -- *Stativité : Typologie et Critères.* Porteuse : M.L.Knittel-F.Namer ATILF-STL-SFL 20 000 euros.
- Projets Loria/ Région Grand-Est/Université de Lorraine SLAM (2015-2018) -- *Schizophrénie et Langage : Analyse et Modélisation*. Porteur: Maxime Amblard. 15 000€
- Projet OrthoCorpus (2015-2017) -- *Construction et exploitation d'un corpus d'articles d'une revue orthophonique*. Porteur : F. Brin-Henry, 25000 euros. ATILF, Région Lorraine, FNO, Ortho-Edition, CH Bar-le-Duc.
- Projet MSH Paris-Saclay HistorIA (2017-2018) -- *Explorer les changements sonores avec l'intelligence artificielle*, Porteuse : Ioana Vasilescu, LIMSI CNRS, Univ. Paris-Saclay. 5000 euros.
- Projet CNRS PEPS HuMaIn (2013-2014) -- *TranSem: Transfert sémantique pour la traduction automatique entre le français et l'allemand*. Porteur : Berthold Crysmann, Membres : Claire Gardent, Yannick Parmentier, Budget : 7.500 €
- Projet interne Labex EFL (2012-2019) -- *ResHau : Ressources linguistique pour le haoussa.* Porteur: Berthold Crysmann, budget 60.000€ environ.
- Projet GD2 Labex EFL (2015-2019) -- *The Syntax of Complex Sentence in Creole Languages*. Porteurs : Stefano Manfredi and Nicolas Quint (40.000€).
- Projet valorisation interne au Labex EFL (2013) -- *TranSem++*. Porteur : Berthold Crysmann, budget: 3.000€
- Projet interne du LabEx EFL (2017-2018) -- *Publication des corpus de textes annotés en cinq langues mandé*. Porteur : V. Vydrin, 5000 euros.
- Projet « Langues et numérique » Ministère de la culture (2017-2018) -- *Production LUdique de Ressources Annotées pour les Langues de France* (PLURAL). Porteur : B. Guillaume (Loria), participants : K. Fort, A. Thibault et A. Millour (Sorbonne Université) et D. Bernhard (LiLPa). 35.000 euros.
- TransAtlantic Platform (2017-2020) -- *Analyzing Children's Language Experiences across the World*. Porteuse France: A. Cristia, 180,000 eurosProjet franco-soudanais (2017) -- *Napata Sudan as a linguistic area*. Porteurs : Maha Aldawi, Nicolas Quint (7000€).
- Projet interne INALCO (2015-2017) -- *MANTAL*, *Traitements et ressources pour les langues mandingues*. Porteur : Damien Nouvel, 18000 euros. ERTIM (D. Nouvel) — LLACAN (V.Vydrin).

Publications récentes des membres du GDR

Abeillé A. and Hassamal S. (2017) Sluicing in Mauritian: a constraint-based approach, CSSP.
Abrusan, M, N. Asher and T. van de Cruys (to appear) 'Meaning shift and grammaticality' In G. Sagi and J. Woods (eds.) The Semantic Conception of Logic. Cambridge University Press
Ackerman, F. and O. Bonami (2017). "Systemic polyfunctionality and morphology-syntax interdependencies." In *Defaults in Morphological Theory*, edited by Andrew Hippisley and Nikolas Gisborne. Oxford: Oxford University Press,
Adamou E. 2016. *A Corpus–driven Approach to Language Contact. Endangered Languages in a Comparative Perspective*. Boston and Berlin: Mouton de Gruyter.
Adamou E. 2017. Subject preference in Ixcatec relative clauses. *Studies in Language* 41 (4): 874-914.

Adda, G., Boula de Mareüil, Ph., Quint, N. and Sichel-Bazin, R. (2017) Norme et variation à l'âge des corpus informatisés pour les langues régionales de France, in : *Usage, norme et codification, de la diversité des situations à l'utilisation du numérique* (dir. Colette Feuillard), Louvain-la-Neuve (Belgique) : EME éditions (ISBN 2806635799), pp. 217-222.

Alxatib, S., I. Gould and O. Percus. Editors for _Snippets_. A journal of technical notes in natural language syntax and semantics. **http://www.lededizioni.it/snippets.html.** LED ON LINE (Electronic Archive of Academic and Literary Texts). ISSN 1590-1807.

An A. and Abeillé A. 2017, Agreement and interpretation of French binomials, in S Müller (ed) Proceedings HPSG Conference, CSLI on-line Publications, p.26-43.

Asher, N., Van de Cruys, T., Bride, A. and Abrusan, M. (2016) 'Integrating type theory and distributional semantics: a case study on adjective-noun compositions'. Computational Linguistics, 42(4), pp. 703-725, doi: 10.1162/COLI_a_00264

Asher, N, M. Abrusan and T. van de Cruys (2016) 'Types, meanings and co-composition in lexical semantics' In S. Chatzikyriakidis and Z. Luo (eds.). Modern Perspectives in Type Theoretical Semantics. Studies of Linguistics and Philosophy, Springer.

Asher, N., Paul S. and Venant A. (2017). Message Exchange Games. *Journal of Philosophical Logic*, vol 46.4, pp.355-404, doi:10.1007/s10992- 016-9402-1.

Auguste J., Rey A. and Favre B. (2017), *"Evaluation of word embeddings against cognitive processes: primed reaction times in lexical decision and naming tasks"*, Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP, 2017

Barra-Jover, M (P. Sauzet, P. col.) 2012 L'évolution des marques du pluriel nominal roman à la lumière de l'occitan in Mario Barra-Jover *et alii*, *Études de linguistique gallo-romane* Presses universitaires de Vincennes, 201-216

Beniamine, S., O. Bonami and B. Sagot. "Inferring Inflection Classes with Description Length." *Journal of Language Modelling* (Sous presse).

Besacier, L., E. Barnard, A. Karpov and T. Schultz. Automatic speech recognition for under-resourced languages: A survey. Speech Communication Journal, Elsevier, vol. 56. 2014.

Billami M.B., Camacho-Collados J., Jacquey E. and Kister E. (*2014*). Annotation sémantique et validation terminologique en texte intégral en SHS. *TALN 2014*, Marseille, 1-4 Juillet.

Bird, S. (2018). Designing mobile applications for documenting endangered languages. In Ken Rehg and Lyle Campbell (eds), Oxford Handbook of Endangered Languages.

Braud C., Lacroix O. and Søgaard A., Does syntax help discourse segmentation? Not so much. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*. Copenhagen (Denmark). Pages 2432-2442. September 2017.

Braud C., Coavoux M. and Søgaard A., Cross-lingual RST Discourse Parsing. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL 2017)*. Valencia (Spain). Pages 292-304. April 2017.

Brin-Henry, **F.**, **Jacquey**, **E., Ollinger, S**. (à paraître). Les termes dits "stratégiques" dans un corpus d'articles scientifiques en orthophonie. LEXIS (11).

Brin-Henry, **F**., **and Knittel, M. L.** (2016). Étude lexicosémantique du nom difficulté (s) dans les comptes rendus de bilan orthophonique: apports structuraux et conceptuels. *Lidil. Revue de linguistique et de didactique des langues*, (53), 19-41.

Brin-Henry, F. (2014).Using Corpus-Based Analyses in Specialised Paramedical French. *Revue Française de Linguistique Appliquée : Langues de spécialité: problèmes et méthodes* (XIX-1), 103-15.

Brun-Trigaud, G., Gaillard-Corvaglia, A ., Léonard, J.-.L, Darlu, P. 2014 Exploration cladistique de l'ALLOc (Atlas Linguistique du Languedoc Occidental), in C. Alén Garbato, C. Torreilles et M.-J.Verny eds *Los que fan viure e treslusir l'occitan*, Actes du Xe Congrès de l'AIEO, Besièrs 12-19 juin 2011), Lemòtges : Lambert-Lucas, 437-448/963.

Buccola, Brian and Spector, Benjamin (2016). 'Modified Numerals and Maximality', *Linguistics and Philosophy 39(3)*: 151-199, doi: 10.1007/s10988-016-9187-2.

Burnett, H. (2017). Sociolinguistic Interaction and Identity Construction: The View from Game Theoretic Pragmatics. *Journal of Sociolinguistics.* 22:238-271.

Burnett, H. (2016). *Gradability in Natural Language: Logical and Grammatical Foundations.* Oxford: Oxford University Press.

Candito, M. and M. Constant. Strategies for Multiword Expression Analysis and Dependency Parsing. In p*roceedings of* the *52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*. Baltimore (United States). Pages 743-753. June 2014.

Caudal, P. (2015). 'Uses of the passé composé in Old French: evolution or revolution ?'. In J. Guéron (ed.), *Sentence and Discourse*, 178–205. Oxford: Oxford University Press.

Caudal, P., H. Burnett and M. Troberg (2016).' Les facteurs de choix de l'auxiliaire en ancien français : étude quantitative'. In S. Prévost and B. Fagard (éds.), *Le français en diachronie. Dépendances syntaxiques, Morphosyntaxe verbale, Grammaticalisation*, 237–265. Bern: Peter Lang.

Chitoran, I., Vasilescu, I., Vieru, B. and Lamel, L. Connected speech in Romanian: exploring sound change through an ASR system, in Recasens D. and Miret, F.S. (eds.), Production and Perception Mechanisms of Sound Change, Lincom Europa Munich, in press.

Coavoux M., Crabbé B. (2017) Incremental Discontinuous Phrase Structure Parsing with the GAP transition, *Proceedings of the 15th Conference of the European Chapter for Computational Linguistics* (EACL) 2017.

Coavoux M., Crabbé B. (2017) Multilingual Lexicalized Constituency Parsing with word-level auxiliary tasks, *Proceedings of the 15th Conference of the European Chapter for Computational Linguistics* (EACL) 2017.

Comorovski, I. (to appear). 'Partitives', in Matthewson L., C. Meier, H. Rullmann, and T. E. Zimmermann (eds.), *The Wiley-Blackwell Companion to Semantics*.

Comorovski, I. (2015/2017**).** 'Focalisation et liage des pronoms : une analyse des pronoms complexes du français et du roumain', in M. S. Istrate et D. Rautu (eds.), *Lucrarile celui de-al saselea simpozion international de lingvistica, Bucureşti, 29–30 mai 2015,* Ed. Univers Enciclopedic Gold, 477-487.

Constant, M., G. Eryiğit, J. Monti, L. van der Plas, C. Ramisch, M. Rosner and A. Todariscu. Multiword Expression Processing: A Survey. In *Computational Linguistics*. 43:4, Pages. 837–892. December 2017.

Constant, M. and J. Nivre. A Transition-based System for Joint Lexical and Syntactic Analysis. In *proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*. Berlin (Germany). Pages 161-171. August 2016.

Crabbé, B., D. Duchier, C. Gardent, J. Leroux and Y. Parmentier. *XMG, eXtensible Meta Grammar.* In Computational Linguistics, 39:3, Pages 581-620 - Septembre 2013.

Cristia, A., Dupoux, E., Gurven, M. and Stieglitz, J. (2017). Child-directed speech is infrequent in a forager-farmer population: A time allocation study. Child Development, 10.1111/cdev.12974.

Crysmann, B. (2015a) Representing morphological tone in a computational grammar of Hausa. Journal of Language Modelling, 3(2), 463–512, 2015.

Crysmann, B. (2015b) **"**Resumption and Extraction in an Implemented HPSG of Hausa". Proceedings of the Grammar Engineering Across Frameworks (GEAF) 2015 Workshop, Beijing, China, July, 65–72, Association for Computational Linguistics, 2015.

Crysmann, B. (2017) "Reduplication in a computational HPSG of Hausa". Morphology, 27(4), 527–561, 2017.

Crysmann, B. and O. Bonami. (2016) "Variable morphotactics in Information-Based Morphology." *Journal of Linguistics* 52: 311-374.

Dal, G. and Namer, F. (to appear). "Playful nonce-formations in French, creativity and productivity". Expanding the Lexicon. Linguistic Innovation, Morphological Productivity, and Ludicity. Arndt-Lappe, S., Braun, A., Moulin, C. and Winter-Froemel, E. Berlin and Boston, De Gruyter. The Dynamics of Wordplay 5.

Duchier, D., Magnana Ekoukou, B., Parmentier, Y., Petitjean, S. and E. Schang. Describing Morphologically-rich Languages using Metagrammars: a Look at Verbs in Ikota. In Workshop on "Language technology for normalisation of less-resourced languages", 8th SALTMIL Workshop on Minority Languages and the 4th workshop on African Language Technology, pages 55–60, Istanbul, Turkey, May 2012b. URL https://hal.archives-ouvertes.fr/hal-00688643. Available on-line at http://aflat.org/files/saltmil8-aflat2012.pdf.

Esher, L. (2016). Morphomic distribution of augments in varieties of Occitan. *Revue Romane*, 51:2, 271-306. doi:10.1075/rro.51.2.09esh

Esher, L. (2017). Morphome death and transfiguration in the history of French. *Journal of Linguistics*, 53:1, 51-84. doi:10.1017/S0022226715000468

Eshkol-Taravella I., Grabar N. Paraphrastic reformulations in spoken corpora. POLTAL 2014. *Advances in Natural Language Processing Lecture Notes in Computer Science.* Volume 8686, p. 425-437, 2014

Florio, S. and Nicolas, D. (2015). Plural logic and sensitivity to order. Australasian Journal of Philosophy 93(3), 444-464.

Fort, K. Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects. 196 p. Wiley-ISTE. Juillet 2016.

Fradin, B. 2017. The multifaceted nature of Denominal Adjectives. Word Structure 10.27-53.

Fradin, B. 2018 (to appear). Competition in derivation: what can we learn from French doublets in *-age* and *-ment*? Competition in morphology, ed. by F. Gardani, F. Rainer, H.C. Luschützky and W.U. Dressler, 000-00. Berlin: Springer.

Gader, N., Ollinger, S. and A. Polguère. *One Lexicon, Two Structures: So What Gives?* Processings of the Seventh Global Wordnet Conference (GWC2014), Jan 2014, Tartu, Estonia. Global WordNet Association, pp.163-171, 2014.

Gardent, C. and L. Perez-Beltrachini. *A Statistical, Grammar-Based Approach to Micro-Planning.* In Computational Linguistics, 43:1, Pages 1-30 - March 2017.

Gardent, C. and S. Narayan *Multiple Adjunction in Feature-Based Tree-Adjoining Grammar* In Computational Linguistics, 41:1, Pages 41-70- March 2015.

Gauthier, E., L. Besacier, S. Voisin. Machine Assisted Analysis of Vowel Length Contrasts in Wolof. Interspeech 2017. Stockholm (Sweden). August 2017.

Giannakidou, A. and A. Mari, (2017). A unified analysis of the future as epistemic modality : the view from Greek and Italian. *Natural Language and Linguistic theory*. https://doi.org/10.1007/s11049-017-9366-z

Grabar N. and Eshkol-Taravella I. (2016). Detection of reformulations in Spoken French. In *proceedings of* LREC2016, 2016

Grabar N. and Eshkol-Taravella I. (2016). Why do we reformulate ? Automatic prediction of pragmatic functions. In *proceedings of HrTAL2016*, 2016

Guillaume, B., K. Fort and N. Lefebvre. Crowdsourcing Complex Language Resources: Playing to Annotate Dependency Syntax. *Proceedings of the International Conference on Computational Linguistics (COLING)*, Osaka, Japon, 2016

Hathout, N. and Namer, F. (à paraître). "La parasynthèse à travers les modèles : des RCL au ParaDis". The lexeme in descriptive and theoretical morphology. Bonami, O., Boyé, G., Dal, G., Giraudo, H.I.n. and Namer, F., Language science Press.

Haug, D. T. T. and T. Nikitina. 2016. Sharing features in agreement. *Natural Language and Linguistic Theory* 34(3): 865-910.

Ingrosso, F., Polguère A. (2015) How Terms Meet in Small-World Lexical Networks: The Case of Chemistry Terminology. *Proceedings of the 11th International Conference on Terminology and Artificial Intelligence (TIA 2015)*, Granada, 167–171.

Iordăchioaia, G., A. Alexiadou and A. Pairamidis. 2017. Morphosyntactic sources for nominal synthetic compounds in English and Greek. In J. Meibauer and P. M. Vogel, eds., *Zusammenbildungen/ Synthetic Compounds*, special issue of *Zeitschrift für Wortbildung/ Journal of Word-formation* 1: 47-72.

Iordăchioaia, G., L. van der Plas and G. Jagfeld. 2016. The grammar of English deverbal compounds and their meaning. In E. Hajicova and I. Boguslavsky (eds.), *Proceedings of the Workshop on Grammar and Lexicon: Interactions and Interfaces* (within COLING 2016), 81–91, Osaka, Japan.

Iordăchioaia, G. and E. Soare. 2015. Pluractionality with lexically cumulative verbs. The supine nominalization in Romanian. *Natural Language Semantics* 23.4: 307-352.

Jacquey, E. and Knittel, M.L. (2015). Nominalisations et Corpus. Verbum XXXVII-1.

Jacquey, E., Kister, L., Marcon, M. and Barreaux, S. 2018. Termes complexes et langues de spécialité en sciences humaines et sociales : que nous apprennent les textes intégraux ?. *Meta 63(1)*.

Knittel, M.L. (2016). Les noms déverbaux : des pluriels lexicaux à la pluriactionnalité. *Lingvisticae Investigationes 39-2 Lexical Plurals and Beyond,* sous la dir. de P. Lauwers and M. Lammert, 373-390.

Knittel, M.L. (2016). A propos de l'indéfinitude des noms d'événements complexes. *Journal of French Language Studies 26-3,* 251-278.

Kruszewski, G., D. Paperno, R. Bernardi and M. Baroni. There is no logical negation here, but there are alternatives. Computational Linguistics 42(4): 637-660

Lev-Ari, S. and Peperkamp, S. (2017). Language for $200: Success in the environment influences grammatical alignment. Journal of Language Evolution, 2, 177-187.

Magri, G. To appear. 'Idempotency, output-drivenness and the faithfulness triangle inequality: some consequences of McCarthy's (2003) categoricity generalization.' To appear in the *Journal of Logic, Language, and Information*.

Maldonado, M, Chemla, E and Spector, Benjamin (2017), Priming plural ambiguities, in *Journal of Memory and Language*, 95:89-101, https://doi.org/10.1016/j.jml.2017.02.002

Mari, A. 2014. Each other, asymmetry and reasonable futures. *Journal of Semantics, 31.2* : 209-261.

Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuko, R., Dupoux, E., and Cristia, A. (2015). Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis. Psychological Science, 26(3), 341-347.

Martin, A., Peperkamp, S. and Dupoux, E. (2013). Learning phonemes with a proto-lexicon. Cognitive Science, 37, 103-124.

Michailovsky, B., M. Mazaudon, A. Michaud, S. Guillaume, A. François and E. Adamou. 2014. Documenting and Researching Endangered Languages: The Pangloss Collection. Language Documentation and Conservation 8 (2014), 119–135.

Michaud, A., S. Guillaume, G. Jacques, D.-K. Mac, M. Jacobson et al. 2016. Contribuer au progrès solidaire des recherches et de la documentation : la Collection Pangloss et la Collection AuCo. In Journées d'Etude de la Parole 2016, Jul 2016, Paris, France. 1, pp.155-163, 2016, Actes de la conférence conjointe JEP-TALN-RECITAL 2016, volume 1 : Journées d'Etude de la Parole.

Mirroshandel G., Nasr A., 2016, *"Integrating Selectional Constraints and Subcategorization Frames in a Dependency Parser"* Computational Linguistics 42(1), 2016

Moot, R. (2015) A type-logical treebank for French, Journal of Language Modelling 3(1), 229-264

Moot, R. (2017) The Grail theorem prover: Type theory for syntax and semantics, In S. Chatzikyriakidis and Z. Luo (eds.). Modern Perspectives in Type Theoretical Semantics. Studies of Linguistics and Philosophy, Springer.

Namer, F., Hathout, N. and Lignon, S. (2017). "Adding morpho-phonological features into a French morpho-semantic resource: the Demonette derivational database". Proceedings of the First International Workshop on Resources and Tools for Derivational Morphology (DeriMo), Milan, Italy: 49-61.

Nasr A., C. Ramisch, J. Deulofeu and A. Valli (2015). Joint Dependency Parsing and Multiword Expression Tokenization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China.

Nicolas, D. (2017). Matière et mélanges. Le Français Moderne 2, 246-260.

Nicolas, D. (2016). Interprétons-nous de la même manière les expressions 'deux pommes' et 'deux pommes et demie'? Travaux de Linguistique 72, 107-119.

Nikitina, T. Forthc. Diminutives derived from terms for children: Comparative evidence from Southeastern Mande. To appear in *Linguistics*.

Nikitina, T. Forthc. Focus marking and differential argument marking: The emergence of bidirectional case marking in Wan. Evangelia Adamou, Katharina Haude and Martine Vanhove (eds.) Information Structure in Lesser-Described Languages. Amsterdam: John Benjamins.

Oliviéri. M, Sauzet P. 2016 « Southern Galloromance : Occitan » in Ledgeway and Maiden eds *The Oxford Guide to the Romance Languages*, 319-349.

Ploux S. (2016) "Lexical semantics and topological models". In Sylvain Loiseau and Jacqueline Léon, editors, *History of Quantitative Linguistics in France*. RAM-Verlag, 2016

Paperno, D., M. Marelli, K. Tentori and M. Baroni. 2014. Corpus-based estimates of word association predict biases in judgment of word co-occurrence likelihood. Cognitive Psychology 74: 66-83.

Polguère A. (2015) Lexicon Embedded Syntax. *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, Uppsala, 2–9.

Polguère A. (2014) From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography* **27(4)**, 396–418.

Qian S., de Groote, P., Amblard, M. (2016) Modal Subordination in Type Theoretic Dynamic Logic, Linguistic Issues in Language Technology, Modes of Modality in NLP, 14(1), 1-54.

Quint, N. (2015) Capeverdean words derived from Portuguese non-infinitive verbal forms: a descriptive and comparative study, in : *Papia* (Revista Brasileira de Estudos Crioulos e Similares), N°25/2, 2015, pp. 189-215.

Quint, N. Forthc. An assessment of the Arabic lexical contribution to contemporary spoken Koalib, in : *Arabic in Contact*, Stefano Manfredi and Mauro Tosco (eds), Amsterdam / Philadelphy : Benjamins (coll. Arabic linguistics).

Renwick, M.E.L., Vasilescu, I., Dutrey, C., Lamel, L., Vieru, B. 2016. Marginal contrast among Romanian vowels : evidence from ASR and functional load **In Proceedings of Interspeech, San Francisco, USA.**

Ribeyre, C., E. de La Clergerie, D. Seddah, *Because Syntax does Matter: Improving Predicate-Argument Structures Parsing Using Syntactic Features*, Proceedings of the Conference of the North-American Chapter of the Association for Computational Linguistics (NAACL 2015), Seattle, USA

Ryzhova, D., M. Kyuseva, and D. Paperno. 2016. Typology of adjectives benchmark for compositional distributional models. Proceedings of the 10th Language Resources and Evaluation Conference, 2016: 1253-1257.

Sauzet, P. 2011 Los morfèmas de plural nominal a Sant Júlian de Crensa (ALLOC 24-03) : [-w] e lo ton bas (Plural morphemes in Saint Julien de Crempse dialect : [-w] and low tone.) in Angelica Rieger ed. *Actes du 9e congrès de l'AIEO* (Aix-la-Chapelle / Aachen / Aisgran 24-31 août 2008), (vol. 2) 827-842.

Sauzet, P., Brun-Trigaud G. 2012 Structure syllabique et évolutions phonologiques en occitan in Mario Barra-Jover *et alii*, *Études de linguistique gallo-romane* Presses universitaires de Vincennes, 161-181.

Sauzet P. 2012 *Occitan Plurals: A Case For A Morpheme Based Morphology* in Hinzelin and Gaglia eds *in* Sascha Gaglia and Marc-Olivier Hinzelin eds Inflection and word formation in Romance languages, Amsterdam ; Philadelphia : John Benjamins Pub. Co., 179-200/ vi-400 p.

Sauzet P. 2012 Geografia linguistica e etimologia : *sens* e *sans* en occitan in Michèle Oliviéri, Guylaine Brun-Trigaud and Philippe Del Giudice eds La leçon des dialectes : Hommage à Jean-Philippe Dalbera, Alessandria : dell'Orso, 337-350 /397 p.

Sauzet P., Brun-Trigaud G. 2014 L'*aso* e l'*òmi* una paradòxa de cronologia relativa en gascon, in C.Alén Garbato, C.Torreilles et M.-J.Verny eds *Los que fan viure e treslusir l'occitan*, Actes du Xe Congrès de l'AIEO, Besièrs 12-19 juin 2011), Lemòtges : Lambert-Lucas, 486-505/963.

Sauzet P., Brun-Trigaud G. 2013 Le Thesaurus Occitan : entre atlas et dictionnaires. *Corpus* 12, 105-140.

Sauzet P. 2016 « Jules Ronjat : la syntaxe et la langue occitane » in Escudé ed. *Autour des travaux de Jules Ronjat, 1913-2013.Unité et diversité des langues*, 43-64 .

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., Quasemizadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., Doucet, A. (2017): "The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions", in the Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)**,** 4 April 2017, Valencia, Spain.

Schang, E., J.-L. Rougé, I. Eshkol, and M. Petit. CreolData : une base de données lexicales sur les langues créoles. Revue Française de Linguistique Appliquée, X(1):65–76, 2005. URL https://halshs.archives-ouvertes.fr/halshs-00764791.

Schang, E. Extended Projections in a Guadeloupean TAG Grammar. In *Workshop on High-level Methodologies for Grammar Engineering@ ESSLLI 2013* (p. 49).

Seddah, D., B. Sagot, M. Candito, V. Mouilleron, V. Combet, *T*he French Social Media Bank: a Treebank of Noisy User Generated Content, COLING 2012, Mumbay, India

Seddah, D., R. Tsarfaty, S. Kübler, M. Candito, J. D. Choi, R. Farkas, J. Foster, I. Goenaga, K. Gojenola Galletebeitia, Y. Goldberg, S. Green, N. Habash, M. Kuhlmann, W. Maier, Y. Marton, J. Nivre, A. Przepiórkowski, R. Roth, W. Seeker, Y.

Spector, Benjamin and Yasutada Sudo, 2017, 'Presupposed Ignorance and Exhaustification: How scalar implicatures and presuppositions interact', *Linguistics and Philosophy 40(5)*: 473-517, doi: 10.1007/s10988-017-9208-9.

Seminck, O. and Amsili P. (2017), *"A Computational Model of Human Preferences for Pronoun Resolution"*, In Proceedings of the SR Workshop at the 15th Conference of the EACL. Valencia, Spain, pp. 53-63.

Seminck, O. and Amsili P. (2018), *A Gold Anaphora Annotation Layer on an Eye Movement Corpus*, Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki (Japan), may 2018.

Tafforeau J,, Bechet F., Artiere T. and Favre, B., 2016, *"Joint syntactic and semantic analysis with a multitask Deep Learning Framework for Spoken Language Understanding"* Interspeech, San Francisco (USA) September 2016 http://pageperso.lif.univ-mrs.fr/~benoit.favre/papers/favre_is2016b.pdf

Tellier I., Eshkol I., Taalab S., Prost J-P. « POS-tagging for Oral Texts with CRF and Category Decomposition », *Research in Computer Science, special issue : Natural Language Processing and its Applications*, p.79-90, 2010.

Trione J., Favre B. and Béchet F. *"Beyond utterance extraction: summary recombination for speech summarization"* Interspeech, San Francisco (USA) September (2016) http://pageperso.lif.univ-mrs.fr/~benoit.favre/papers/favre_is2016a.pdf

Versley, V. Vincze, M. Woliski, A. Wróblewska, E. Villemonte de la Clergerie, *Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically Rich Languages,* 2013, Proceedings of the Fourth SPMRL Workshop, Seattle, USA.

Vydrin, V.. New Electronic Resources for Texts in Manding Languages. In: Daniela Merolla and Mark Turin (eds.). Searching For Sharing: Heritage and Multimedia in Africa. Cambridge, UK: Open Book Publishers, 2017, pp. 109-121.

Vydrin, V.. Quantifiers in Dan-Gwɛɛtaa (South Mande). In: Denis Paperno, Edward L. Keenan (eds.). Handbook of quantifiers in natural language: Volume II. [Studies in linguistics and philosophy, vol. 97]. Springer, 2017, pp. 203-280.

Vydrin, V. and Rovenchak, A. and Maslinsky, K.. Maninka Reference Corpus: A Presentation. In: TALAf 2016 : Traitement automatique des langues africaines (écrit et parole) Atelier JEP-TALN-RECITAL 2016 - Paris le 4 juillet 2016. Actes. http://talaf.imag.fr/2016/Actes/VYDRIN_ET_AL%20-%20Maninka%20Reference%20Corpus:%20A%20Presentation.pdf

Wang, R., Zhao, H., Ploux, S., Lu, B. L., and Utiyama, M. (2016)**.** A Bilingual Graph-Based Semantic Model for Statistical Machine Translation. In *IJCAI* (pp. 2950-2956).

Warlaumont, A., vanDam, M., Bergelson, E. and Cristia, A. (2017). HomeBank: A repository for long-form real-world audio recordings of children. Proceedings of Interspeech Show and Tell.

Zanon Boito, M., Berard, A., Villavicencio, A. and L. Besacier. Unwritten Languages Demand Attention Too! Word Discovery with Encoder-Decoder Models. IEEE ASRU 2017. Okinawa (Japan). December 2017.

Financement

Budget Annuel demandé: 30 KEuros
Financement partiel de l'école d'été: 20KE

Frais de mission pour déplacements internes: 6 KEuros
Mobilité des doctorants: 10 KEuros
Invitation de scientifiques étrangers: 10 KEuros
Vacations pour le gestionnaire du site WEB: 4 KEuros

Contacts

claire.gardent@loria.fr
denis.paperno@loria.fr