Nous considérons dans notre travail la tâche du traitement automatique visant à construire, à partir de textes issus d'un corpus de constats d'accidents de la route, des interprétations compatibles avec ces derniers, et à en proposer des illustrations sous forme de séquences d'images fixes. Notre recherche est le fruit d'une collaboration entre un laboratoire universitaire et une entreprise. Elle prend appui sur le modèle de la Grammaire Applicative et Cognitive [DES 90], qui vise en particulier à "expliquer", à un certain niveau cognitif, les transferts entre représentations imagées et verbales. Pour une revue de la question relative à la "transcription automatique Verbal-Image", nous renvoyons à [ARN 90] ; et plus particulièrement aux travaux de C. Vandeloise [VAN 87] et du groupe "Langue, Raisonnement, Calcul" de l'Université Paul Sabatier [AUR 90, SAB 95] ainsi qu'aux approches proposées dans [ARN 93] et dans le système SPRINT [YAM 92]. Plus proches encore de nos préoccupations, B. Victorri et P. Enjalbert [ENJ 94, POI 95] posent le problème de l'animation visuelle issue de l'interprétation de textes. Nous présentons dans cet article, à travers le traitement d'un exemple, la méthode générale d'analyse que nous avons adoptée, qui s'appuie en priorité sur des connaissances linguistiques. Le texte pris comme exemple est le suivant : Je roulais sur la partie droite de la chaussée quand un véhicule arrivant dans le virage a été complètement déporté. Serrant à droite au maximum, je n'ai [pu] éviter la voiture [qui arrivait à grande vitesse]. Nous ne traitons pas ici la modalité introduite par pu , de même que la relative qui arrivait à grande vitesse. Dans la première partie de l'article, nous présentons l'architecture globale du système informatique. Dans la deuxième partie, nous proposons des éléments d'analyse pour une solution opératoire aux problèmes d'articulation des significations lexicales et grammaticales, sous forme d'une segmentation du texte en différentes phases spatio-temporelles. Dans la troisième partie, nous présentons une modélisation des lieux de circulation et du mouvement des véhicules garantissant le passage à l'image.

Nous donnons ici un aperçu du logiciel DECID développé au GETA afin d'informatiser le processus de rédaction du dictionnaire explicatif et combinatoire du français contemporain.

Diverses méthodes ont été proposées pour construire un "graphe conceptuel" représentant le "sens" d'une phrase à partir de son analyse syntaxique. Cependant, peu d'entre elles reposent sur un véritable formalisme linguistique. Nous nous intéressons ici à la construction d'une telle représentation sémantique à partir de la représentation syntaxique produite par une analyse LFG, et montrons comment une transposition du "joint dirigé" des graphes conceptuels permet d'effectuer cette construction à partir de la "structure sémantique "des LFG.

Le terme de lambda-DRT désigne un ensemble de méthodes permettant de construire des représentations sémantiques (DRS) à partir d'arbres syntaxiques. La mise en oeuvre de telles méthodes nécessite l'élaboration de systèmes de types dont le détail est rarement présenté. C'est à la description d'un tel système que cet article est consacré.

Dans cet article, nous comparons deux modèles linguistiques utilisés en TAL, les grammaires d'arbres adjoints [= TAG] et le Théorie Sens-Texte [= TST]. Nous montrons que ces deux modèles présentent des similitudes notables, et que les représentations les plus abstraites qu'ils donnent d'une phrase ? la représentation sémantique en TST et l'arbre de dérivation en TAG ? sont équivalentes. De ce rapprochement découle d'une part que l'on peut s'inspirer de la procédure de dérivation TAG pour opérer la correspondance Sens-Texte, et d'autre part que l'on peut concevoir une grammaire TAG comme le résultat de la précompilation d'une grammaire Sens-Texte.

Dans le cadre des approches à base de grammaires faiblement sensibles au contexte, cette contribution passe en revue le problème de l'extraction de l'arbre d'analyse le plus probable dans le modèle du Data-Oriented Parsing (DOP). Une démonstration formelle de l'utilisabilité des méthodes de Monte-Carlo est donnée, puis une technique d'échantillonnage contrôlée est développée permettant de garantir (avec un certain seuil de confiance fixé a priori) que l'arbre d'analyse

sélectionné est bien l'arbre d'analyse le plus probable au sens de DOP. plus probable au sens de DOP.

Dans cet article, nous proposons de prendre du recul par rapport à l'aspect opératoire du tagging, et nous tentons de montrer que le tagging ouvre la voie au renouveau de l'analyse syntaxique en la fondant sur l'explicitation des processus: processus de déduction locale dans les syntagmes non récursifs, et processus de mise en relation des syntagmes non récursifs, étendant ainsi à l'analyse syntaxique les propriétés calculatoires du tagging.

Nous décrivons dans cet article un système d'extraction automatique de réponses. L'extraction automatique de réponses (EAR) a pour but de trouver les passages d'un document qui répondent directement à une question posée par un utilisateur. L'EAR est plus ambitieuse que la recherche d'informations et l'extraction d'informations dans le sens que les résultats de la recherche sont des phrases et non pas des documents en entier, et dans le sens que les questions peuvent être formulées de façon libre. Elle est par contre moins ambitieuse que les systèmes questions-réponses car les réponses ne sont pas générées à partir d'une base de connaissance, mais repérées dans les textes des documents. La version actuelle d'ExtrAns permet d'analyser la documentation en ligne (en Anglais) du système Unix (les "man pages"), et de construire une représentation sémantique (sous forme d'expressions logiques) des phrases. Un programme de démonstration de théorèmes trouve ensuite les passages pertinents qui sont mis en évidence dans leur contexte.

L'analyseur syntaxique robuste que nous décrivons dans cet article s'insère dans le cadre des travaux relatifs au traitement du langage oral. Nous montrons à partir d'une étude menée sur des dialogues transcrits qu'il est avantageux de traiter le langage oral avec un analyseur syntaxique robuste faisant appel à un analyseur syntaxique conçu pour l'écrit. Pour ce faire, nous avons réalisé un système qui se base sur une architecture à deux couches : le noyau et la périphérie, l'une

interagissant avec l'autre via un superviseur. La première couche, le noyau, est dédiée à l'analyse des constituants d'énoncés respectant la grammaire standard : c'est un analyseur syntaxique de l'écrit. La deuxième couche, la périphérie, se charge de "corriger" les constituants de l'énoncé oral ayant subis des distorsions. Cette couche intervient lorsque le noyau ne parvient plus à progresser dans son analyse. L'ordre d'intervention de ces deux couches dans le processus d'analyse est déterminé par l'occurrence de marques de surface qui signalent la présence d'une distorsion ou d'une construction particulière (interrogative, relative, etc.). Le système ainsi conçu nous a permis de traiter les bruits et différents types de répétitions caractéristiques de l'oral. La première version de l'analyseur nous a fourni des résultats encourageants qui nous permettent de confirmer l'interdépendance entre le traitement du langage oral et le traitement du langage écrit. Ces résultats nous invitent aussi à reconsidérer le rôle et l'utilité, pour le traitement de l'oral, des ressources d'informatique linguistique développées pour l'écrit.

Les sorties parlées d'un système de dialogue homme-machine sont souvent d'une qualité médiocre pour deux raisons : le texte généré par le générateur ne tient souvent pas compte de l'usager et la synthèse de la parole à partir de ce texte, donne souvent une prosodie de lecture. L'article décrit les principes d'introduction d'une composante pragmatique en amont du système de génération afin de tenir compte des degrés de force et des buts illocutoires. En effet, les études en génération portent généralement sur les réalisations linguistiques du contenu propositionnel de la réponse (c'est-à-dire le sujet du message) mais très rarement sur la constitution de l'acte énonciatif qui est généralement abandonnée au moteur de dialogue de l'application. Nous développons cet aspect au coeur même de la génération afin de fournir un outil performant et surtout pour prendre en compte les aspects pragmatiques de manière précoce dans le dialogue. La conclusion est que notre système, grâce à son module de traitement pragmatique, a la capacité de générer des énoncés plus naturels. La génération est basée non plus sur le contenu informatif seul, mais aussi sur une situation pragmatique, exprimée dans un cadre bien défini. Notre approche montre l'intérêt de prendre en

compte les paramètres illocutoires dans le cadre du dialogue.

Nous présentons ici le dispositif GASPAR qui construit des représentations des mots sous la forme d'objets informatiques appelés des prototypes ; GASPAR associe à ces objets les comportements syntaxiques et sémantiques des mots en prenant appui sur des informations extraites à partir d'un corpus. GASPAR a pour première tâche de construire progressivement une représentation informatique des mots, sans présumer de leurs descriptions linguistiques ; il doit ensuite reclasser les mots représentés et mettre au jour, de manière inductive, les classes de mots du sous-langage étudié. Nous montrons comment la programmation à prototypes permet de représenter des mots dynamiquement par apprentissage et par affinements successifs. Elle permet ensuite d'amorcer un début de classement de ces mots sur la base de leurs contraintes syntaxico-sémantiques en construisant des hiérarchies locales de comportements partagés.

Se déplacer à pied ou par un autre moyen est une activité humaine courante. Cependant, trouver son chemin suppose souvent des aides de type verbal (descriptions d'itinéraires) ou iconique (croquis, cartes). Nous présentons dans cet article un système capable de produire des descriptions d'itinéraires en métro. Notre générateur est fondé sur un mod èle cognitif de la production de telles descriptions et sur une analyse de corpus. Nous montrons comment des facteurs comme l'importance relative d'informations et les choix stylistiques peuvent faire varier à la fois le contenu et la forme de la description d'un itinéraire.

Les outils d'aide à la construction de terminologie à partir de corpus ont connu un essor important ces dernières années. D'un autre côté, les outils d'aide à l'acquisition de relations sémantiques entre termes sont peu nombreux. Face à ce problème, nous avons développé le système Prométhée qui acquière incrémentalement un ensemble de patrons lexico-syntaxiques caractéristiques d'une relation sémantique. Ainsi, pour la relation d'hyponymie, Prométhée a extrait

un ensemble de patrons qui servent à améliorer la couverture d'un thesaurus ou d'une base de connaissances.

Cet article présente une chaîne de traitement automatique réalisée dans le cadre du projet ILIAD (Informatique Linguistique et Infométrie pour l'Analyse de grands fonds Docu-mentaires) du GIS Sciences de la Cognition. Cette cha?ine est dédiée à l'analyse de l'infor-mation à partir de corpus de textes de très grand volume, en français. Elle est expérimentée sur un corpus de 2,5 Mb et a conduit à la création de 50 classes de termes. Ces classes sont construites sur la base de la co-occurrence des termes et représentent des connaissances du do-maine. Les différentes étapes de la cha?ine associent des méthodes linguistiques informatiques et des méthodes statistiques : pré-traitement des textes, étiquetage, morphologie, terminologie et analyse des documents. Pour chacune d'entre elles, nous présentons les méthodes, les outils ainsi que leur evaluation.

Dans le domaine de l'ingénierie linguistique et de la connaissance, le problème des ressources lexicales et linguistiques s'est toujours posé. Néanmoins, l'avancée des techniques du Traitement Automatique des Langues Naturelles (TALN) l'a rendu plus sensible. Il nous faut maintenant pouvoir répondre à des besoins importants en terme de quantité, de qualité et de complexité. La complexité et la diversité des informations requises augmente avec les exigences des outils de TALN ainsi qu'avec le développement de nouvelles applications (humaines ou machinales). Si la récupération (semi)automatique d'information lexicale est une piste, elle ne pourra remplacer la création manuelle de dictionnaires. Nous nous sommes donc intéressés à la construction d'outils pour lexicographes et lexicologues. pour répondre aux besoins de nos systèmes de traduction et à la demande du projet Universal Networking Language (UNL), nous avons décidé d'informatiser la construction d'une base lexicale multi-lingue. Dans ce but, nous avons fusionné des dictionnaires existants. A partir de ces données, nous générons automatiquement des fichiers qui sont envoyés aux lexicographes. Ceux-ci complètent et corrigent les données sur leur plate-forme avec des outils

très simples. Les fichiers sont ensuite réintégrés dans la base lexicale. La dernière étape est la génération de dictionnaires nécessaires à nos systèmes de traduction.

Le traitement de la sémantique de la spatialité nécessite de manipuler des entités qu'il n'est pas aisé de définir. On parle souvent de lieux, mais on les assimile trop à un sens hors contexte des mots désignant des objets ; on évoque les notions de routine et de trajectoire, mais elles ne sont pas réellement intégrées au calcul du sens. Nous discutons dans cet article de la façon d'intégrer ces référents spatiaux à un traitement automatique, et nous focalisons en particulier sur la richesse que recèlent les chemins et les trajectoires dans l'expression du déplacement lorsqu'un formalisme leur donne existence.

L'expression de la spatio-temporalité est traditionnellement scindée en deux paradigmes, la localisation et le déplacement. La localisation exprime alors un certain nombre de relations entre une entité à localiser et des sites, tandis que le déplacement exprime un changement de ces relations dans le temps. Pourtant, c'est omettre l'autonomie et la richesse du déplacement que de l'exprimer par rapport à la localisation, et c'est aussi opposer deux paradigmes qui partagent un certain nombre de types de contraintes (topologie, distance, etc.). Nous proposons donc d'observer le domaine spatio-temporel et ses articulations d'une façon qui ne les oppose pas mais qui montre au contraire ce qu'ils partagent. Ce travail de formalisation et d'analyse est destiné à l'élaboration de mécanismes de compréhension automatique.

Cet article présente la démarche suivie pour mettre en place un logiciel d'aide à la veille technologique. Pour l'analyse de documents très techniques, les veilleurs utilisent des outils d'infométrie, qui sont pertinents sur les données structurées, mais qui ne sont pas adaptés pour l'exploitation des informations textuelles. Nous avons donc réalisé un logiciel d'extraction d'informations, nommé VIGITEXT. Notre approche, basée sur la définition de notions indépendantes

du domaine comme l'/amélioration/, l'/augmentation/ ou l'/utilisation/, permet d'extraire des informations textuelles à partir d'abrégés descriptifs de brevets rédigés en anglais sans utiliser de lexique technique ou de calculs statistiques. De plus, cette méthode est opérationnelle pour tous les sujets de veille, et les résultats, qui sont les extraits organisés selon les notions, sont simples à utiliser par des veilleurs. Dans cet article, nous décrivons les particularités de la veille technologique, et les limites des logiciels généralement utilisés. Ensuite, nous détaillons l'exploitation de notions générales basée sur la définition de connaissances linguistiques et qui met en oeuvre la méthode d'exploration contextuelle. Nous présentons enfin le prototype VIGITEXT, avec ses spécificités et ses utilisations possibles dans une démarche de veille.

L'action GRACE est le premier exemple d'application du paradigme d'évaluation aux étiqueteurs morpho-syntaxiques pour le français dans le cadre d'une campagne d'évaluation formelle, à participation ouverte et utilisant des données de grande taille. Après une rapide description de l'organisation et du déroulement de l'action ainsi que des problèmes posés par la nécessaire mise en place d'un référentiel commun pour l'évaluation, nous présenterons en détail la métrique Précision-Décision qui a été développée dans le cadre de GRACE pour la mesure quantitative des performances des systèmes d'étiquetage. Nous nous intéresserons ensuite aux résultats obtenus pour les participants à la phase de test de la campagne et indiquerons les aspects du protocole d'évaluation qui restent encore à valider sur les données recueillies. Enfin, nous conclurons en soulignant les incidences positives d'une campagne d'évaluation comme GRACE sur le domaine de l'ingénierie linguistique.

L'objectif de cette étude concerne le traitement d'homophones singulier/pluriel dans un Système de Reconnaissance de la Parole en exploitant les contraintes d'accord dans la phrase à reconnaître. Un certain nombre de ces contraintes ne peut être traité par les modèles de langage à portée locale de type n-gram utilisés habituellement. Les deux modèles proposés, le modèle à base de syntagme

et le modèle Homophone-Cache, permettent de résoudre certains cas d'homophonie par deux méthodes différentes : le modèle à base de syntagme permet d'introduire des contraintes syntaxiques ; le modèle Homophone-Cache a pour objet de discriminer les homophones singulier/pluriel, de manière robuste, en étant peu sensible à la mauvaise reconnaissance d'un mot au sein de la phrase.

Nous présentons dans cet article une nouvelle approche, que nous appelons 5P, permettant la description des propriétés d'un langage et son utilisation pour une analyse automatique. Nous montrons comment cette approche permet la prise en compte de la dimension descriptive de la linguistique. Par ailleurs, nous présentons une technique d'analyse, appelée analyse par Filtrage et Fusion, qui tire parti de cette description en propriétés. Nous montrons en quoi ces deux projets (description d'une langue et analyse automatique) convergent et ouvrent de nouvelles perspectives.

Une des recherches de pointe menée actuellement en informatique est l'extraction des connaissances dans un texte électronique (textual data mining). Ce thème de recherche est de première importance pour les technologies de l'information qui sont confrontées à des marées de documents électroniques. Pour résoudre ce problème, plusieurs stratégies sont possibles : les unes relèvent des mathématiques et les autres de l'informatique linguistique. Nous présentons dans cet article un modèle hybride, à la fois robuste et fin, qui s'inspire des modèles neuronaux et de l'analyse linguistique informatique.

L'interrogation de bases de données en langue naturelle est une application directe du traitement automatique des langues naturelles. Son utilité va en s'accroissant avec le développement d'outils d'information accessibles au grand public à travers la Toile Internet. L'approche que nous proposons s'appuie d'une part sur les fondations linguistiques établies par la théorie de Z. S. Harris (dans l'élaboration du dictionnaire, et surtout dans la définition des opérateurs linguistiques), et d'autre part

sur un outil informatique précis (les transducteurs). Elle représente une alternative aux traitements syntaxico-sémantiques habituellement développés dans des formalismes logiques. Elle s'appuie sur la constitution d'une bibliothèque d'opérateurs linguistiques pour les domaines d'application.

Cet article présente l'identification en corpus des adjectifs relationnels considérés par les linguistes comme hautement dénominatifs. Notre approche utilise un programme d'extraction terminologique qui s'applique sur un corpus préalablement étiqueté et lemmatisé. Après avoir rappelé quelques propriétés linguistiques des adjectifs relationnels, nous présenterons le programme d'extraction de terminologie et les modifications apportées à celui-ci pour effectuer cette identification. Nous évaluerons le caractère dénominatif de ces adjectifs et des termes nominaux où ils apparaissent en les comparant à un thesaurus. Nous conclurons sur l'intérêt de ces adjectifs à la fois pour l'extraction de terminologie mais aussi pour d'autres problématiques comme l'extraction de connaissances à partir de corpus ou la mise à jour d'un thesaurus.

Cette communication décrit un outil informatique de construction et de consultation d'un lexique verbal saisi sur des supports informatiques en vue d'une utilisation par des linguistes et qui peut être appelé à certaines étapes d'un traitement automatique de textes écrits. L'analyse du lexique verbal s'inscrit dans un modèle, celui de la Grammaire Applicative et Cognitive (GAC) développé dans l'équipe LaLIC. Le formalisme utilisé est celui du ?-calcul typé et de la logique combinatoire typée avec ses combinateurs. Le lexique verbal est organisé à l'aide d'un langage de représentation sémantico-cognitif (LRSC) s'appuyant sur un ensemble de relateurs et de primitives sémantico-cognitives typées. Dans un premier temps nous présentons un outil informatique (DISCC) qui a pour tâche d'aider un sémanticien à construire des représentations sémantico-cognitives associées aux significations des verbes; et dans un second temps, nous montrons comment il est possible de consulter les différentes significations d'un vocable verbal polysémique représenté sous forme d'un réseau. La présentation ne présente pas un dictionnaire

mais développe une méthodologie de construction et de manipulation d'une base de connaissances sémantico-cognitives des verbes.

Dans ce bref document, nous présentons des résultats préliminaires d'une méthode de description de la sémantique des formes prédicatives dans un cadre génératif. Nous proposons une méthode pour découper les sens, en identifiant les types d'inférences qu'ils entrainent. Nous proposons une analyse intégrée des métaphores et des métonymies, ainsi qu'une représentation des sens sous forme typée et sous-spécifiée.

En synthèse automatique de la parole, la phonétisation est une étape cruciale pour une bonne intelligibilité et une bonne qualité de voix. Elle consiste à convertir une suite de mots en chaîne phonétique, qui sera par la suite utilisée pour générer le signal sonore. Les homographes hétérophones et les ajustements phonologiques tels que la liaison et l'élision sont les sources d'erreurs les plus courantes. De plus, des mots comme 'plus' , 'tous' et certains nombres ('cinq', 'six', 'dix',?) pour lesquels plusieurs réalisations phonétiques sont possibles, peuvent également être problématiques. Nous proposons ici une résolution de ces cas complexes par l'utilisation d'une analyse syntaxique.

La morphologie médicale est riche et productive. À côté de la simple flexion, dérivation et composition sont d'autres moyens pour créer des mots nouveaux. La connaissance morphologique se révèle par conséquent très importante pour toute application dans le traitement automatique du langage médical. Nous proposons une méthode simple et puissante pour l'acquisition automatique d'une telle connaissance. Cette méthode tire avantage de listes de termes synonymes disponibles afin d'amorcer le processus d'acquisition. Nous l'avons expérimentée dans le domaine médical sur le Microglossaire de Pathologie SNOMED. Les familles de mots morphologiquement reliés que nous avons obtenues sont correctes à 95 %.Utilisées dans un outil d'aide au codage avec expansion de

requête, elles permettent d'en améliorer les performances.

Le traitement automatique du langage requiert des corpus textuels de plus en plus volumineux, entre autres pour les étiqueteurs morpho-syntaxiques. Ces processus de traitement ne sont pas exempts d'erreurs. Dans l'optique d'améliorer cet étiquetage de corpus hétérogènes (composés de textes tout-venant), une approche adaptative au type de texte utilisant les ressources produites par une campagne d'évaluation sera proposée. Les résultats d'une première validation seront présentés sur les données MULTITAG. Les faits suivants sont constatés : les textes ne sont pas homogènes en terme de distribution de parties du discours, les classifications a priori ne fournissent pas une homogénéité en terme de performance et un même texte peut produire des variations positives pour un système et négatives pour un autre. De plus, il existe une relation entre la typologie de textes obtenue de façon non supervisée sur le jeu de caractères et les variations de performance.

Depuis [Kimball 73], les préférences d'attachement telles que "l'association droite" et "l'attachement minimal" ont essentiellement été formulées en termes d'arbres de constituants (e.g. forme, nombre de noeuds ...) . Nous présentons 2 principes de préférence d'attachement formulés en termes d'arbres de dérivation (i.e. d'information dépendancielle) dans le cadre du formalisme des Grammaires d'Arbres Adjoints Lexicalisées (LTAG) . Nous montrons pourquoi ce type d'approche permet de remédier aux défauts des approches structurales exprimées en termes d'arbres de constituants et rendent compte d'heuristiques largement acceptées (i.e. argument / modifieur, idiomes).

Nous nous intéressons ici aux méthodes d'alignement automatique destinées à produire des corpus bi-textuels, utiles au traducteur, au terminologue ou au linguistique. Certaines techniques ont obtenu des résultats probants en s'appuyant sur la détermination empirique des « cognats » (de l'anglais « cognate »), des mots qui se traduisent l'un par l'autre et qui présentent une ressemblance

graphique. Or les cognats sont généralement captés au moyen d'une approximation abrupte, de nature opératoire : on considère tous les 4-grammes (mots possédants 4 lettres en commun) comme cognats potentiels. Aucune étude n'a été faite, à notre connaissance, à propos de la validité de cette approximation. Afin d'en démontrer les possibilités et les limites, nous avons cherché à déterminer empiriquement la qualité de cette simplification, en termes de bruit et de silence (ou de manière complémentaire, de précision et de rappel). Nous avons ensuite essayé de développer un filtrage plus efficace, basé sur l'utilisation des sous-chaînes maximales. Enfin, nous avons corrélé les améliorations du filtrage avec les résultats de l'alignement, en nous basant sur une méthode générale développée par nous : nous avons pu constater un net progrès en terme de rappel et de précision de l'alignement.

Nous présentons une technique de résolution de proformes enchâssées à l'aide des méta-structures Prolog. Nous montrons tout d'abord un exemple d'utilisation de ces méta-structures pour contrôler l'appartenance d'un élément à un domaine. Une plus grande utilité est ensuite démontrée dans la résolution de contraintes contextuelles dynamiques, qui sont particulières dans le sens où elles interviennent en fonction des contraintes déjà existantes sur les éléments considérés. Une application utile de ces contraintes est d'éviter les redondances dans la recherche des possibilités de référents pour un discours considéré, notamment dans le cas de proformes enchâssées.

Nous présentons un système dédié à la conception et au test d'un sous-language d'application pour un système de Dialogue Homme-Machine. EGAL se base sur une grammaire LTAG générale de la langue qui est spécialisée à une application donnée à l'aide d'un corpus d'entraînement. Un double effort a porté premièrement sur la définition d'une méthodologie précise passant par une expérimentation de type Magicien d'Oz pour le recueil des corpus et des estimations de la représentativité du corpus de conception, et, deuxièmement, sur la spécification des composants du

système en vue de mettre en oeuvre des outils convivaux, génériques et ouverts.

L'article décrit l'implémentation d'un modèle d'intonation dans son application à la synthèse de la parole pour le français. Le modèle se caractérise par l'importance accordée à la syntaxe et par une approche analytique de l'intonation qui, en synthèse, permet une manipulation explicite et compositionnelle du sens intonatif. Le traitement proprement dit est précédé d'une analyse syntaxique identifiant les constituants, certains rapports de dépendance ou certaines constructions qui demandent une intonation particulière. Ces aspects intonatifs sont représentés par des marqueurs symboliques. À partir de l'arborescence sont constitués les groupes intonatifs, tout en tenant compte du rythme. Dans certaines conditions, des réajustements de la structure syntaxique seront effectués. Les tons mélodiques sont attribués aux groupes en fonction des marqueurs et des rapports syntaxiques.

Cet article a pour but de décrire la mise au point et l'expérimentation de méthodes d'apprentissage de syntaxe à partir d'exemples positifs, en particulier pour des applications de Reconnaissance de la Parole et de Dialogue Oral. Les modèles syntaxiques, destinés à être intégrés dans une chaîne de traitement de la parole, sont extraits des données par des méthodes d'inférence grammaticale symbolique et stochastique. Ils sont fondés sur des techniques de correction d'erreurs dans les séquences. L'ensemble de ce travail a été réalisé dans le cadre du contrat 97- 1B-004 avec France-Telecom (Centre National d'Etudes des Télécommunications). Dans la première partie de cet article, nous rappellons les distances entre séquences basées sur des opérations élémentaires de correction d'erreur. Nous décrivons ensuite un algorithme classique d'inférence grammaticale fondé sur cette notion, et nous en proposons une amélioration. Nous abordons à cet endroit le problème de l'évaluation d'un concept appris seulement à partir d'exemples positifs, sans contre-exemples. Par la suite, le modèle syntaxique est étendu en attribuant des probabilités (apprises à partir des données) aux règles de la grammaire. On dispose dans ce cadre d'un outil

d'évaluation de la qualité de l'apprentissage : la perplexité ; cependant pour obtenir des résultats significatifs, il faut être capable de probabiliser l'espace entier des séquences, ce qui implique de lisser la grammaire stochastique apprise. Une technique de lissage est proposée, qui permet alors d'évaluer l'apprentissage sur le corpus de données issues de l'expérimentation en dialogue oral.

Cet article présente les avantages qu'apporte la modélisation des ressources linguistiques utilisées dans une application. Le lecteur trouvera également dans cet article une présentation rapide de deux méthodes répandues dans le monde de l'informatique (Merise et UML) et leur modèle associé (entité relation et objet). Enfin, nous donnerons un exemple de modélisation des ressources linguistiques d'une application en cours de développement.

Quels types d'informations sont nécessaires à l'interprétation de référents évolutifs et de référents associés ? Nous verrons que les anaphores évolutives et associatives sont construites à partir de processus et de situations, et que leur interprétation nécessite une représentation lexicale complexe. Les approches atomiques peuvent par conséquent difficilement rendre compte de ce type d'anaphores : cependant les propriétés des quantificateurs semblent jouer un rôle dans ces phénomènes.

Dans cet article, nous montrons, à travers l'exposé de résultats d'une expérience menée sur corpus, comment la connaissance des thèmes dans lesquels apparaissent des mots et la mise en évidence de similarités et de différences entre les voisinages de leurs occurrences dans les parties de textes abordant ces thèmes permettent de mettre au jour des différences fines dans les acceptions associées aux mots dans chacun de ces thèmes. La méthode proposée pour ce faire est presque entièrement automatique et est basée sur le calcul d'intersections et de différences ensemblistes entre des séquences de mots constituant des contextes.

A des fins d'automatisation de la vérification de traduction, les méthodes traditionnelles se basent généralement sur un fort niveau de littéralité dans le style de la traduction. En faisant appel à des bases terminologiques multi-lingues et des algorithmes d'alignement de textes parallèles, il est possible de vérifier dans un travail de traduction le respect de normes strictes, sous la forme d'une liste de possibilités de traduction pour un terme donné. Nous proposons ici une méthode alternative basée sur le repérage, dans les deux textes, de structures sémantiques générales, ou isotopies, et la comparaison des schémas qu'elles présentent au niveau du texte et non plus de la phrase ou du paragraphe, permettant ainsi une plus grande tolérance dans le style de traduction à vérifier.

L'analyse des propositions relatives en anglais telle que déecrite par Sag (1997) se base sur une classification à deux dimensions des constructions syntaxiques en HPSG. Nous présentons ici une implémentation de cette analyse, fondée sur l'héritage multiple et les templates à deux dimensions dans le système ProFIT (Erbach, 1995).

Nous traitons dans ce papier du problème de la détection et de la correction des graphies fautives dans les textes arabes. Nous commençons par présenter une expérience visant à mesurer de manière comparative la difficulté du problème pour l'arabe, le français et l'anglais. L'idée est d'évaluer le degré de "ressemblance" (proximité) des mots au sein de chaque langue. Ensuite les algorithmes de base de notre méthode de correction sont présentés.

Nous présentons dans cet article l'architecture logicielle de Context, plate-forme d'ingénierie linguistique dédiée au filtrage sémantique. Nous avons défini un modèle conceptuel et un langage de description et de traitement des connaissances linguistiques. Ces connaissances sont gérées par un système dédié et indépendant des applications qui les utilisent. Les traitements sont spécifiés sous forme déclarative dans un langage formel que nous présentons.

Nous présentons les principaux résultats obtenus à ce jour dans le cadre du projet MAREDI qui vise à développer un système de traitement de la langue naturelle permettant d'analyser des transcriptions de dialogues oraux et de générer un modèle conceptuel de la conversation. Nous discutons principalement des aspects touchant l'analyse sémantique, en l'occurrence le rôle qu'y jouent les actes de discours et l'analyse casuelle, tout en présentant brièvement l'architecture globale du système et les caractéristiques de ses différentes composantes.

Cet article introduit une représentation du sens basée sur des vecteurs de notions. Ces vecteurs sémantiques ont pour but de rendre compte de lÕensemble des idées évoquées dans un segment textuel. Ce type de représentation utilisé en conjonction avec une analyse morpho-syntaxique classique permet dÕeffectuer dans de nombreux cas une désambiguïsation lexicale efficace.

Nous présentons dans cet article une approche acquisitionniste de la langue naturelle dans le cadre du dialogue homme-machine finalisé, ainsi qu'une première implémentation. Le système de dialogue COALA, qui tente de mettre en oeuvre cette approche, se constitue ses propres représentations à partir de son expérience, au lieu d'utiliser des connaissances langagières pré-codées. Les premiers dialogues obtenus sont de ce fait laborieux, mais ils deviennent progressivement conviviaux. Le système COALA réunit d'une part un modèle de dialogue (non décrit dans cet article) et d'autre part une méthode d'analyse par chart hypothético-déductif qui permet une forme d'apprentissage par extraction de régularités structurelles.

Notre étude propose un modèle cohérent pour formaliser la modalité pour une implémentation en tant que module interlingua d'un système de traduction automatique (TA). Un grand nombre d'erreurs de traduction en TA peut être attribué à l'absence d'un traitement autonome de modalité. Le modèle tient compte de l'hétérogénéité des éléments modaux et permet la combinaison des éléments déclencheurs.

Nous présentons une méthode de classification d'analyses robustes sur des hypothèses concurrentes d'un système de reconnaissance de la parole. Pour réaliser cette classification, différents critères hétérogènes sont combinés, comme le score de reconnaissance, diverses caractéristiques syntaxiques et sémantiques propres à l'analyse robuste effectuée ou encore des estimations de la cohérence pragmatique. L'analyse est fondée sur une variante des LTAG (Lexicalized Tree Adjoining Grammars). La classification proposée est évaluée à partir d'un corpus d'analyses robustes d'hypothèses de reconnaissance.

Nous présentons une technique d'analyse robuste dans le but de relayer la décision d'un système de reconnaissance de la parole. La stratégie d'analyse proposée est fondée sur une grammaire d'arbres adjoints lexicalisée compactée et sur la mise en concurrence des différentes hypothèses du système de reconnaissance de la parole. Les problèmes de robustesse sont étudiés en considérant les interférences entre erreurs de reconnaissance de la parole et phénomènes de parole spontanée dans les dialogues homme-machine.

Cet article décrit l'amorçage d'une ontologie et d'un lexique partagé dans une population d'agents robotiques dotés de capacités visuelles. Cette évolution a lieu alors que les agents jouent un jeu de langage, appelé "guessing game". Nous étudions les dynamiques d'un tel système et montrons, en particulier, comment la synonymie et l'ambiguïté du système sémantique, qui émergent dans un premier temps, sont progressivement réduites au fur et à mesure que l'environnement physique se complexifie.

L'analyse syntaxique robuste est devenue une technique essentielle à toute application qui touche au contenu des documents. Les analyseurs inscrits dans cette approche permettent d'extraire des informations d'ordre linguistique qui peuvent être exploitées postérieurement par des traitements

linguistiques plus profonds ou par des systèmes de recherche d'information. Une des caractéristiques principales de ces outils est leur robustesse. Or, cette robustesse est souvent diminuée par la grande hétérogénéité de phénomènes linguistiques et extralinguistiques présents dans les textes tout-venant. Cet article présente tout d'abord (section 1) la notion de robustesse et caractérise (section 2) les systèmes d'analyse syntaxique robuste. L'article présente par la suite (section 3) un inventaire de phénomènes linguistiques et extralinguistiques non-standard attestés dans divers corpus et, finalement, (section 4) une architecture qui se propose de traiter ces phénomènes.

Un des enjeux de la CHM orale, dès qu'elle aura quitté le champ d'investigation du dialogue fortement finalisé, semble être de pouvoir allier robustesse (face aux spécificités de l'oral), efficacité et couverture de la langue. Cet article tente de montrer qu'une analyse linguistique détaillée peut être menée tout en respectant la contrainte de robustesse imposée. Le système, proposé comme une alternative aux méthodes sélectives, repose tout d'abord sur l'exploitation du pouvoir structurant de la syntaxe au niveau de constituants minimaux non récursifs (chunks). Une recherche des relations de dépendances entre les têtes lexicales associées à ces unités peut être ensuite envisagée à un niveau sémantico-pragmatique.

Cet article présente un système de compréhension d'une requête orale dans le cadre d'un dialogue homme-machine. Si le dialogue est finalisé, son domaine est néanmoins relativement vaste : la construction d'une représentation sémantique de l'énoncé exige une analyse linguistique plus fine que celle utilisée dans les applications classiques de CHM1. Le système présenté ici, entièrement lexicalisé, utilise à différents niveaux les types logiques : au niveau syntaxique, il reconstitue les groupes de mots par composition de lambda-termes ; à un niveau plus sémantique, il compose les différents constituants pour construire une formule logique qui correspond à la représentation sémantique de la requête. Les tests réalisés semblent valider la méthode.

Il est communément admis que la tâche de désambiguïsation sémantique n'est pas une fin en soi. Pour tenter d'apporter un début de solution à ce problème reconnu comme très difficile, de nombreux systèmes ont été développés. Pour la plupart, ces systèmes sont destinés à être les composants de systèmes plus complexes (moteurs de recherche d'information, de dialogue personne-machine, ou d'aide à la traduction). Néanmoins, ils sont testés en tant que tels dans le cadre de campagnes d'évaluation, comme par exemple Senseval ou Romanseval. La seconde édition de ces campagnes est d'ores et déjà planifiée. De fait, on est en droit de se demander - sans pour autant vouloir chercher à enrayer le mouvement -, si la désambiguïsation sémantique a un sens, et si oui lequel. Il ne faut pas voir dans ce questionnement un jeu de mots gratuit, mais bien la nécessité de soumettre à l'examen une pratique dans laquelle s'engagent de plus en plus de chercheurs, qu'ils soient linguistes ou informaticiens. Si l'on s'en tient au protocole suivi lors de la première campagne d'évaluation de Senseval, on peut dégager de ses caractéristiques un certain nombre d'observations qui peuvent alimenter la réflexion. Une quarantaine de mots appartenant à l'une ou l'autre de trois catégories grammaticales avait été retenue : les noms, les verbes et les adjectifs. Pour chacun de ces mots était fournie une liste d'étiquettes sémantiques et pour couvrir l'ensemble de ces sens, en moyenne, une centaine d'exemples étiquetés ainsi qu'une définition pour chaque étiquette. Pour chaque mot, enfin une centaine d'exemples de tests devaient être étiquetés par les différents systèmes en lice. Pour un mot donné, les étiquettes pouvaient entretenir des relations de type hiérarchique, ce qui permettait d'évaluer les systèmes à trois niveaux de granularité : fin, grossier, et intermédiaire. Une remarque préalable concerne le corpus d'apprentissage disponible pour chacun des mots. Pour un mot donné, seul le mot en question était étiqueté. Pour les mots du contexte aucune étiquette sémantique n'était proposée. Les annotations sémantiques posées par des juges humains sur chacun des exemples relatifs à un mot particulier, avait fait l'objet d'un arbitrage, et quand cela s'avérait impossible plusieurs étiquettes sémantiques avaient été maintenues. Enfin, détail qui peut avoir son importance : les étiquettes sémantiques

utilisées pour annoter le corpus d'apprentissage étaient plus fines que celles qui étaient employés pour le niveau le plus fin d'évaluation. Notre propos n'est pas ici de décrire ifficultés à mettre en relation des définitions et des emplois de mots en contexte .Une des significations d'un mot employé dans un contexte particulier peut se trouver absente de la ressource pour plusieurs raisons. Les lacunes des dictionnaires ont suffisamment été pointées du doigt à diverses reprises, pour qu'il soit nécessaire d'en rajouter sur le sujet. Par essence, une ressource finie ne peut couvrir toutes les productions résultant des capacités créatives qui s'exercent sur les langages naturels. Certains usages langagiers correspondent à des nuances fines dont il est difficile de rendre compte dans un lexique où par contre figurent souvent des acceptions qui n'ont plus cours. Par ailleurs, il n'y a pas de découpage unique d'un mot en unités de sens. Il suffit pour s'en convaincre de comparer les choix faits par différents dictionnaires. Mais, le problème est plus complexe que cela. En analysant le fonctionnement des métaphores, on peut expliquer comment certaines figures de style permettent de rajouter un sens (le plus souvent figuré) à un mot tout en maintenant en partie son sens premier. Ces évidences expliquent en grande partie la complexité de la relation entre étiquetage et choix d'étiquettes sémantiques. Les méthodes numériques ont leur mot à dire pour tenter de trouver une voie entre lexique et corpus annoté. Toute approche qui entre dans cette catégorie peut non seulement permettre de choisir une étiquette parmi plusieurs, mais aussi servir à classer toutes les étiquettes candidates soit par calcul de distances ou de vraisemblances. Si la méthode retenue est de ce type, le vecteur final associé à un exemple peut être vu comme un moyen de localiser un emploi particulier dans l'espace déterminé par la base que forment les étiquettes sémantiques. Par le biais d'une analyse en composantes principales ou d'une analyse discriminante, des axes orthogonaux peuvent être dégagés un à un, axes correspondant à un compromis entre le jeu d'étiquettes initial et les exemples présents dans le corpus annoté. Même si le processus n'a pas tendance à converger, il ne serait peut-être pas inutile de de le voir comme une étape parmi d'autres d'une procédure itérative appliquée s'il le faut sur des données mouvantes afin de reproduire les aspects dynamiques de toute langue vivante. Si l'on accepte l'idée que Numérique et

Métrique ont un rôle à jouer dans le domaine de la Sémantique, il est possible de voir le problème de la désambiguïsation sémantique comme formant un tout avec celui du choix des étiquettes. La question ne serait plus comment choisir entre tel ou tel sens pour un emploi donné, mais dans quelle région se situe cet emploi, sachant que la somme des usages aura tendance à modifier l'espace lui-même, dès qu'il sera patent qu'il aura été pour une raison ou pour autre, sous ou sur dimensionné.

Dans cet article, nous présentons la problématique de l'hétérogénéité des données textuelles et la possibilité d'utiliser cette dernière pour améliorer les traitements automatiques du langage naturel. Cette hypothèse a été abordée dans (Biber, 1993) et a donné lieu à une première vérification empirique dans (Sekine, 1998). Cette vérification a pour limite de ne s'adapter qu'à des textes dont le type est explicitement marqué. Dans le cadre de textes tout venant, nous proposons une méthode pour induire des types de textes, apprendre des traitements spécifiques à ces types puis, de façon itérative, en améliorer les performances.

Le Web ou les bibliothèques numériques offrent la possibilité d'interroger de nombreux serveurs d'information (collections ou moteurs de recherche) soulevant l'épineux problème de la sélection des meilleures sources de documents et de la fusion des résultats provenant de différents serveurs interrogés. Dans cet article, nous présentons un nouvelle approche pour la sélection des collections basée sur les arbres de décision. De plus, nous avons évalué différentes stratégies de fusion et de sélection permettant une meilleure vue d'ensemble des différentes solutions.

Dans cet article, nous décrivons un système de traduction automatique pour l'allemand, le français, l'italien et l'anglais. Nous utilisons la technique classique analyse-transfert-génération. Les phrases d'entrée sont analysées par un analyseur générique multi-lingue basé sur la théorie ((Principes & Paramètres)) de la grammaire générative chomskienne. Le mécanisme de transfert agit sur des

représentations hybrides qui combinent des éléments lexicaux avec de l'information sémantique abstraite. Enfin, un générateur inspiré de la même théorie linguistique engendre des phrases de sortie correctes. Nous décrivons également brièvement les différentes interfaces envisagées sur Internet.

La sémantique de certains verbes (doubler, distancer, suivre) et de certaines prépositions ou adverbes (devant, derrière) peut poser problème dès lors qu'elle est considérée comme purement spatiale, c'est-à-dire en des termes " classiques " comme la topologie, le repérage ou la distance. Nous proposons dans cet article une description plus générale de ces items lexicaux basée sur la notion d'axe abstrait, rendant compte de leur sens dans différents domaines, ainsi que les différents mécanismes permettant de les plonger dans le domaine qui concerne notre recherche, le spatio-temporel. Ces mécanismes sont intégrés dans un modèle informatique de génération automatique de prédicats verbaux afin d'éprouver leur pertinence.

Tous les formalismes linguistiques font usage de la notion de contrainte qui, dans son sens le plus large, indique une propriété devant être satisfaite. Les contraintes sont extrêmement utiles à la fois pour représenter l'information linguistique, mais également pour en contrôler le processus d'analyse. Cependant, l'usage qui est fait des contraintes peut être très différent d'une approche à l'autre : dans certains cas, il s'agit simplement d'un mécanisme d'appoint, dans d'autres, les contraintes sont au coeur de la théorie. Il existe cependant un certain nombre de restrictions à leur utilisation, en particulier pour ce qui concerne leur implantation. Plus précisément, s'il semble naturel (au moins dans certains paradigmes) de considérer l'analyse syntaxique comme un problème de satisfaction de contraintes, on constate cependant qu'il est extrêmement difficile de réaliser concrètement une telle implantation. Ce constat est en fait révélateur d'un problème dépassant le simple cadre de l'implémentation : nous montrons dans cet article qu'une approche totalement basée sur les contraintes (permettant donc de concevoir l'analyse comme un problème de satisfaction) est

incompatible avec une interprétation générative classique accordant un statut particulier à la relation de dominance. Nous proposons ici un cadre permettant à la fois de tirer parti des avantages des grammaires syntagmatiques tout en s'affranchissant des problèmes liés aux approches génératives pour ce qui concerne l'usage des contraintes en tant qu'unique composant grammatical. Nous présentons ici cette approche, les Grammaires de Propriétés, ainsi que leur implémentation.

Dans cet article nous introduisons la notion de grammaire transductive, c'est-à-dire une grammaire formelle définissant une correspondance entre deux familles de structures. L'accent sera mis sur le module syntaxique de la théorie Sens-Texte et sur une famille élémentaire de grammaires de dépendance transductives. Nous nous intéresserons à la comparaison avec les grammaires génératives, ce qui nous amènera à discuter de l'interprétation des modèles génératifs actuels.

Dans cet article, nous montrons comment le paradigme d'évaluation peut servir pour produire de façon plus économique des ressources linguistiques validées de grande qualité. Tous d'abord nous présentons le paradigme d'évaluation et rappelons les points essentiels de son histoire pour le traitement automatique des langues, depuis les premières applications dans le cadre des campagnes d'évaluation américaines organisées par le NIST et le DARPA jusqu'aux derniers efforts européens en la matière. Nous présentons ensuite le principe qui permet de produire à coût réduit des ressources linguistiques validées et de grande qualité à partir des données qui sont produites lorsque l'on applique le paradigme d'évaluation. Ce principe trouve ses origines dans les expériences (Recognizer Output Voting Error Recognition) qui ont été effectuées pendant les campagnes d'évaluation américaine pour la reconnaissance automatique de la parole. Il consiste à combiner les données produites par les systèmes à l'aide d'une simple stratégie de vote pour diminuer le nombre d'erreurs. Nous faisons alors un lien avec les stratégies d'apprentissages automatiques fondées sur la combinaison de systèmes de même nature. Notre propos est illustré par la description de la production du corpus MULTITAG (projet du programme Ingénierie des

Langues des département SPI et SHS du CNRS) à partir des données qui avaient été annotées lors de la campagne d'évaluation GRACE, correspondant à un corpus d'environ 1 million de mots annotés avec un jeu d'étiquettes morpho-syntaxiques de grain très fin dérivé de celui qui a été défini dans les projets EAGLES et MULTEXT. Nous présentons le corpus MULTITAG et la procédure qui a été suivie pour sa production et sa validation. Nous concluons en présentant le gain obtenu par rapport à une méthode classique de validation de marquage morho-syntaxique.

Cet article présente une méthode d'étiquetage sémantique de noms propres fondé sur la technique des arbres de décision. Ces derniers permettent de modéliser les éléments saillants dans les contextes d'occurrence de noms propres d'une classe donnée. Les arbres de décision sont construits automatiquement sur un corpus d'apprentissage étiqueté, ils sont ensuite utilisés pour étiqueter des noms propres apparaissant dans un corpus de test. Les résultats de l'étiquetage du corpus de test est utilisé pour enrichir un lexique de noms propres. Ce dernier peut être utilisé à son tour pour réestimer les paramètres d'un étiqueteur stochastique. Nous nous intéressons en particulier au cas où le corpus de test a été glané sur le Web.

Les bi-textes sont des corpus bilingues parallèles, généralement segmentés et alignés au niveau des phrases. Une des applications les plus directes de ces corpus consiste à en extraire automatiquement des correspondances lexicales, fournissant une information utile aux traducteurs, aux lexicographes comme aux terminologues. Comme pour l'alignement, des méthodes statistiques ont donné de bons résultats dans ce domaine. Nous pensons qu'une exploitation judicieuse d'indices statistiques adaptés et d'algorithmes de conception simple permet d'obtenir des correspondances fiables. Après avoir présenté les indices classiques, auxquels nous essayons d'apporter des améliorations, nous proposons dans cette article une étude empirique destinée à en montrer les potentialités.

Cette étude porte sur l'acquisition des Entités Nommées (EN) à partir du Web. L'application présentée se compose d'un moissonneur de pages et de trois analyseurs surfaciques dédiés à des structures spécifiques. Deux évaluations sont proposées : une évaluation de la productivité des moteurs en fonction des types d'EN et une mesure de la précision.

Cet article présente une application permettant d'écrire des requêtes complexes sur des corpus étiquetés et de formater librement les résultats de ces requêtes. Le formalisme des requêtes est basé sur le principe des expressions régulières bien connu de la plupart des linguistes travaillant sur des corpus écrits. Contrairement à certains logiciels, qui ne permettent que l'extraction de concordances au format relativement figé, le formatage libre du résultat des requêtes permet leur réutilisation par des programmes ultérieurs et autorise une grande diversité d'applications, s'écartant largement du cadre des simples concordanciers.

La quantité de documents disponibles via Internet explose. Cette situation nous incite à rechercher de nouveaux outils de localisation d'information dans des documents et, en particulier, à nous pencher sur l'algorithmique des grammaires context-free appliquée à des familles de graphes d'automates finis (strictement finis ou à cycles). Nous envisageons une nouvelle représentation et de nouveaux traitements informatiques sur ces grammaires, afin d'assurer un accès rapide aux données et un stockage peu coûteux en mémoire.

Cet article présente un système de reconnaissance des noms propres pour le Français. Les spécifications de ce système ont été réalisées à la suite d'une étude en corpus et s'appuient sur des critères graphiques et référentiels. Les critères graphiques permettent de concevoir les traitements à mettre en place pour la délimitation des noms propres et la catégorisation repose sur les critères référentiels. Le système se base sur des règles de grammaire, exploite des lexiques spécialisés et comporte un module d'apprentissage. Les performances atteintes par le système, sur les

anthroponymes, sont de 89,4% pour le rappel et 94,6% pour la précision.

Tous les médias continus (parole, texte, musique, cinéma) ont, par définition, une structure linéaire, à partir de laquelle un processus cognitif est capable de reconstituer une organisation temporelle différente. Mais jusqu'à quel point faut-il comprendre un texte pour le segmenter en situations et les articuler entre elles ? Autrement dit : jusqu'à quel point faut-il connaître la musique pour différencier couplet et refrain ? Dans un grand nombre de cas, il est possible d'effectuer une telle segmentation automatiquement, et cela uniquement à partir d'indices morpho-syntaxiques. Notre prototype de programme identifie des situations référentielles et analyse la façon dont elles sont articulées pour reconstruire la structure temporelle d'un récit. L'objectif de cette communication n'est pas la description de ce programme, mais plutôt le point de vue du linguiste : comment détecter les discontinuités, c'est-à-dire comment décider s'il y a complétion ou rupture.

Dans cet article, nous présentons le système QALC (Question Answering Langage Cognition) qui a participé à la tâche Question Réponse de la conférence d'évaluation TREC. Ce système a pour but d'extraire la réponse à une question d'une grande masse de documents. Afin d'améliorer les résultats de notre système, nous avons réfléchi à la nécessité de développer, dans le module d'analyse, le typage des questions mais aussi d'introduire des connaissances syntaxico-sémantiques pour une meilleure recherche de la réponse.

Un nombre important de requêtes soumises aux moteurs de recherche du W3 ne satisfont pas pleinement les attentes des utilisateurs. La liste de documents proposée en retour est souvent trop longue : son exploration représente un travail exagérément laborieux pour l'auteur de la requête. Nous proposons d'apporter une valeur ajoutée aux systèmes de recherche documentaire (RD) existants en y ajoutant un filtrage n'utilisant que des données fournies par l'utilisateur. L'objectif de notre étude est de confronter un modèle dynamique de la mémoire sémantique des individus (ou

des agents) développé par notre équipe à une tâche nécessitant une compétence interprétative de la part des machines. Nous souhaitons dépasser la sémantique lexicale couramment utilisée dans ce champ d'application pour aboutir à l'utilisation d'une sémantique des textes et accroître par ce biais, à la fois la qualité des résultats et la qualité de leur présentation aux usagers.

Ce papier présente la première partie d'un travail de thèse qui vise à construire un « dictionnaire distributionnel » à partir d'un corpus de référence. Le dictionnaire proposé est basé sur un ensemble de critères différentiels stricts qui constituent des indices exploitables par des machines pour discriminer le sens des mots en contexte. Pour l'instant, le travail a porté sur 50 000 occurrences qui ont été étiquetées de façon manuelle. Ce sous-corpus pourra servir de corpus d'amorçage pour la constitution d'un corpus étiqueté plus grand, qui pourrait servir à différents tests et travaux sur la désambiguïsation automatique.

Les progrès réalisés ces dernières années dans le domaine du traitement automatique des langues naturelles (TALN) ouvrent la voie à des traitements encore plus sophistiqués dans lesquels la sémantique devrait tenir une place centrale. Notre objectif, à long terme, est de réaliser un analyseur texte vers sens s'appuyant sur la théorie Sens-Texte d'Igor Mel'cuk. Cette analyse viserait une compréhension plus approfondie du texte, permettant donc d'atteindre une représentation de niveau sémantique, et une grande robustesse face à des entrées plus ou moins bien formées telles que celles issues de dialogues oraux. Mais renverser la théorie Sens-Texte passe par la définition et la mise en oeuvre de structures de données et d'algorithmes spécifiques pour la représentation et la manipulation automatique des informations linguistiques, notamment des entrées lexicales. Pour cela, nous proposons l'utilisation du paradigme de programmation par contraintes qui offre un moyen efficace d'atteindre nos objectifs.

Nous présentons dans cet article le projet au sein duquel nous développons un logiciel permettant

d'assister l'utilisateur lors de la formulation de sa requête de recherche sur le Web et de personnaliser des sous-ensembles du Web selon ses besoins informationnels. L'architecture du logiciel est basée sur l'intégration de plusieurs outils numériques et linguistiques de traitements des langues naturelles (TALN). Le logiciel utilise une stratégie semi-automatique où la contribution de l'utilisateur assure la concordance entre ses attentes et les résultats obtenus. Ces résultats sont stockés dans diverses bases de données permettant de conserver différents types d'informations (classes de sites/pages Web similaires, profils de l'usager, lexiques, etc.) constituant une projection locale et personnalisée du Web.

L'indexation audiovisuelle, indispensable pour l'archivage et l'exploitation des documents, se révèle être un processus délicat, notamment à cause de la multiplicité de significations qui peuvent être attachées aux images. Nous proposons dans cette communication une méthode d'instanciation de "patrons d'indexation" à partir d'un corpus d'articles de journaux écrits. Cette méthode repose sur un processus "d'amorçage hiérachisé", qui permet de trouver de nouveaux termes à partir de termes connus dans leur voisinage et de leurs relations taxinomiques sous forme d'ontologie.

Les automates et transducteurs pondérés sont utilisés dans un éventail d'applications allant de la reconnaissance et synthèse automatiques de la langue à la biologie informatique. Ils fournissent un cadre commun pour la représentation des composants d'un système complexe, ce qui rend possible l'application d'algorithmes d'optimisation généraux tels que la déterminisation, l'élimination des mots vides, et la minimisation des transducteurs pondérés. Nous donnerons un bref aperçu des progrès récents dans le traitement de la langue à l'aide d'automates et transducteurs pondérés, y compris une vue d'ensemble de la reconnaissance de la parole avec des transducteurs pondérés et des résultats algorithmiques récents dans ce domaine. Nous présenterons également de nouveaux résultats liés à l'approximation des grammaires context-free pondérées et à la reconnaissance à l'aide d'automates pondérés.

Nous proposons de montrer comment l'analyse syntaxique automatique est aujourd'hui à un tournant de son évolution, en mettant l'accent sur l'évolution des modèles d'analyse syntaxique : de l'analyse de langages de programmation (compilation) à l'analyse de langues, et, dans le cadre de l'analyse de langues, de l'analyse combinatoire à l'analyse calculatoire, en passant par le tagging et le chunking (synthèse en section 4). On marquera d'abord le poids historique des grammaires formelles, comme outil de modélisation des langues et des langages formels (section 1), et comment la compilation a été transposée en traduction automatique par Bernard Vauquois. On analysera ensuite pourquoi il n'a pas été possible d'obtenir en analyse de langue un fonctionnement analogue à la compilation, et pourquoi la complexité linéaire de la compilation n'a pas pu être transposée en analyse syntaxique (section 2). Les codes analysés étant fondamentalement différents, et le tagging ayant montré la voie, nous en avons pris acte en abandonnant la compilation transposée : plus de dictionnaire exhaustif en entrée, plus de grammaire formelle pour modéliser les structures linguistiques (section 3). Nous montrerons comment, dans nos analyseurs, nous avons implémenté une solution calculatoire, de complexité linéaire (section 5). Nous conclurons (section 6) en pointant quelques évolutions des tâches de l'analyse syntaxique.

Dans cet article nous présentons les premiers résultats de l'exploitation d'un Corpus français arboré (Abeillé et al., 2001). Le corpus comprend 1 million de mots entièrement annotés et validé pour les parties du discours, la morphologie, les mots composés et les lemmes, et partiellement annotés pour les constituants syntaxiques. Il comprend des extraits de journaux parus entre 1989 et 1993 et écrits par divers auteurs, et couvre différents thèmes (économie, littérature, politique, etc.). Après avoir expliqué comment ce corpus a été construit, et comment l'exploiter à l'aide d'un outil de recherche spécifique, nous exposerons quelques résultats linguistiques concernant les fréquences et les préférences lexicales et syntaxiques. Nous expliquerons pourquoi nous pensons que certains de ces résultats sont pertinents en linguistique théorique et en psycholinguistique.

L'objectif de notre travail est de construire une représentation sémantique d'un corpus de textes français au sein des graphes conceptuels simples. Notre conceptualisation est fondée sur les Schèmes Sémantico-Cognitifs et la théorie aspecto-temporelle introduits par J. P. Desclés. Un texte est représenté par deux structures. La première modélise la représention semanticocognitive des propositions du texte, et la seconde le diagramme temporel exprimant les contraintes temporelles entre les différentes situations décrites dans le texte. La prise en compte de ces deux structures et des liens qu'elles entretiennent nous a amenés à modifier le modèle des graphes conceptuels simples et à envisager les modes d'interaction entre temps, aspect (grammatical) et significations des lexèmes verbaux.

Nous présentons dans cet article un modèle d'exploration contextuelle et une plate-forme logicielle qui permet d'accéder au contenu sémantique des textes et d'en extraire des séquences particulièrement pertinentes. L'objectif est de développer et d'exploiter des ressources linguistiques pour identifier dans les textes, indépendamment des domaines traités, certaines des relations organisatrices des connaissances ainsi que les organisations discursives mises en places par l'auteur. L'analyse sémantique du texte est guidée par le repérage d'indices linguistiques déclencheurs dont l'emploi est représentatif des notions étudiées.

Le sujet du présent article est l'intégration des sens portés par les mots en contexte dans une représentation vectorielle de textes, au moyen d'un modèle probabiliste. La représentation vectorielle considérée est le modèle DSIR, qui étend le modèle vectoriel (VS) standard en tenant compte à la fois des occurrences et des co-occurrences de mots dans les documents. L'intégration des sens dans cette représentation se fait à l'aide d'un modèle de Champ de Markov avec variables cachées, en utilisant une information sémantique dérivée de relations de synonymie extraites d'un dictionnaire de synonymes.

Une véritable classification numérique multi-lingue est impossible si on considère seulement le mot comme unité d'information privilégiée. En traitant les mots comme jetons, la tokenisation s'avère relativement simple pour le français et l'anglais, mais très difficile pour des langues comme l'allemand ou l'arabe. D'autre part, la lemmatisation utilisée comme moyen de normalisation et de réduction du lexique constitue un écueil non moins négligeable. La notion de n-grams, qui depuis une décennie donne de bons résultats dans Pidentification de la langue ou dans l'analyse de l'oral, est, par les recherches récentes, devenue un axe privilégié dans l'acquisition et l'extraction des connaissances dans les textes. Dans cet article, nous présenterons un outil de classification numérique basé sur le concept de n-grams de caractères. Nous évaluons aussi les résultats de cet outil que nous comparons à des résultats obtenus au moyen d'une classification fondée sur des mots.

Cet article propose une description des dépendances à distances s'appuyant sur une approche totalement déclarative, les grammaires de propriétés, décrivant l'information linguistique sous la forme de contraintes. L'approche décrite ici consiste à introduire de façon dynamique en cours d'analyse de nouvelles contraintes, appelées propriétés distantes. Cette notion est illustrée par la description du phénomène des disloquées en français.

En travaillant sur l'interrogation de bases de données en langue naturelle, nous sommes amenés à exploiter les propositions du Laboratoire de Linguistique Informatique (LLI) en matière de représentation de la langue : les classes d'objets. Un outil d'interrogation définit une application du langage vers le modèle de l'information stockée. Ici les classes d'objets et leurs prédicats appropriés modélisent le langage source, tandis que le modèle relationnel sert pour les données interrogées. Nous présentons d'abord ce contexte d'application, puis comment nous utilisons les classes d'objets et prédicats appropriés dans ce cadre.

La transcription manuelle de la prosodie est une tâche extrêmement coûteuse en temps, qui requiert des annotateurs très spécialisés, et qui est sujette à de multiples erreurs et une grande part de subjectivité. Une automatisation complète n'est pas envisageable dans l'état actuel de la technologie, mais nous présentons dans cette communication des outils et une méthodologie qui permettent une réduction substantielle du temps d'intervention manuelle, et améliorent l'objectivité et la cohérence du résultat. De plus, les étapes manuelles nécessaires ne demandent pas une expertise phonétique poussée et peuvent être menées à bien par des étudiants et des "linguistes de corpus".

Trouver l'arbre d'analyse le plus probable dans le cadre du modèle DOP (Data-Oriented Parsing) ? une version probabiliste de grammaire à substitution d'arbres développée par R. Bod (1992) ? est connu pour être un problème NP-difficile dans le cas le plus général (Sima'an, 1996a). Cependant, si l'on introduit des restrictions a priori sur le choix des arbres élémentaires, on peut obtenir des instances particulières de DOP pour lesquelles la recherche de l'arbre d'analyse le plus probable peut être effectuée en un temps polynomial (par rapport à la taille de la phrase à analyser). La présente contribution se propose d'étudier une telle instance polynomiale de DOP, fondée sur le principe de sélection miminale-maximale et d'en évaluer les performances sur deux corpus différents.

La plupart du temps, les études qui portent sur l'aggrégation des phrases en génération de texte, se focalisent sur l'utilisation des connecteurs pour relier les phrases courtes et inventées. Mais, les connecteurs limitent le nombre des unités qu'il est possible de combiner à la fois. Comment condenser l'information en peu d'unités, sans utiliser trop de connecteurs ?.7 Cette étude porte sur des documents ayant trait à la biologie et discute de l'agrégation des phrases par les auteurs quand ils résument. Cet article présente aussi quelques préalables et difficultés pour un système de

résumé automatique. Beaucoup de phrases sont aggrégées sans signe explicite, ni connecteur, ni ponctuation.

Nous présentons dans cet article le système QALC qui a participé à la tâche Question Answering de la conférence d'évaluation TREC. Ce système repose sur un ensemble de modules de Traitement Automatique des Langues (TAL) intervenant essentiellement en aval d'un moteur de recherche opérant sur un vaste ensemble de documents : typage des questions, reconnaissance des entités nommées, extraction et reconnaissance de termes, simples et complexes, et de leurs variantes. Ces traitements permettent soit de mieux sélectionner ces documents, soit de décider quelles sont les phrases susceptibles de contenir la réponse à une question.

Afin d'améliorer les performances des systèmes de résumé automatique ou de filtrage sémantique concernant la prise en charge de la cohérence thématique, nous proposons un modèle faisant collaborer une méthode d'analyse statistique qui identifie les ruptures thématiques avec un système d'analyse linguistique qui identifie les cadres de discours.

Les nombreuses recherches portant sur le phénomène de la liaison en français ont pu mettre en évidence l'influence de divers paramètres linguistiques et para-linguistiques sur la réalisation des liaisons. Notre contribution vise à déterminer la contribution relative de certains de ces facteurs en tirant parti d'une méthodologie robuste ainsi que d'outils de traitement automatique du langage. A partir d'un corpus de 5h de parole produit par 10 locuteurs, nous étudions les effets du style de parole (lecture oralisée/parole spontanée), du débit de parole (lecture normale/rapide), ainsi que la contribution de facteurs syntaxiques et lexicaux (longueur et fréquence lexicale) sur la réalisation de la liaison. Les résultats montrent que si plusieurs facteurs étudiés prédisent certaines liaisons, ces facteurs sont souvent interdépendants et ne permettent pas de modéliser avec exactitude la réalisation des liaisons.

Cet article décrit une cascade de transducteurs pour l'extraction de noms propres dans des textes. Après une phase de pré-traitement (découpage du texte en phrases, étiquetage à l'aide de dictionnaires), une série de transducteurs sont appliqués les uns après les autres sur le texte et permettent de repérer, dans les contextes gauches et droits des éléments "déclencheurs" qui signalent la présence d'un nom de personne. Une évaluation sur un corpus journalistique (journal Le Monde) fait apparaître un taux de précision de 98,7% pour un taux de rappel de 91,9%.

Cet article présente un nouvel algorithme de détection de motifs syntaxiques récurrents dans les textes écrits en langage naturel. Il décrit d'abord l'algorithme d'extraction fondé sur un modèle d'édition généralisé à des arbres stratifiés ordonnés (ASO). Il décrit ensuite les expérimentations qui valident l'approche préconisée sur des textes de la littérature française classique des XVIIIe et XIXe siècle. Une sous-partie est consacrée à l'évaluation empirique de la complexité algorithmique. La dernière sous-partie donnera quelques exemples de motifs récurrents typiques d'un auteur du XVIIIe siècle, Madame de Lafayette.

Nous présentons dans cet article un système de Compréhension Automatique de la Parole (CAP) tentant de concilier les contraintes antinomiques de robustesse et d'analyse détaillée de la parole spontanée. Dans une première partie, nous montrons l'importance de la mise en oeuvre d'une CAP fine dans l'optique d'une Communication Homme-Machine (CHM) sur des tâches moyennement complexes. Nous présentons ensuite l'architecture de notre système qui repose sur une analyse en deux étapes : une première étape d'analyse syntaxique de surface (Shallow Parsing) générique suivie d'une seconde étape d'analyse sémantico-pragmatique ? dépendante du domaine d'application ? de la structure profonde de l'Énoncé complet.

Le processus de construction de terminologie ne peut être entièrement automatisé. Les méthodes

et des outils de la terminologie computationnelle permettent de prendre en charge une partie de la tâche, mais l'expertise humaine garde une place prépondérant. Le défi pour les outils terminologiques est de dégrossir les tâches qui sont soit trop longues soit trop complexes pour l'utilisateur tout en permettant à ce dernier d'intégrer ses propres connaissances spécialisées et en lui laissant le contrôle sur la terminologie à construire. Nous montrons ici comment le rôle de cette expertise est pris en compte dans SynoTerm, l'outil d'acquisition de relation de synonymie entre termes que nous avons dÉveloppé.

Cet article présente une méthode de construction automatique de liens morphologiques à partir d'un dictionnaire de synonymes. Une analyse de ces liens met en lumière certains aspects de la structure morphologique du lexique dont on peut tirer partie pour identifier les variations allomorphiques des suffixations extraites.

La synonymie est une relation importante en TAL mais qui reste problématique. La distinction entre synonymie relative et synonymie subjective permet de contourner certaines difficultés. Dans le cadre des vecteurs conceptuels, il est alors possible de définir formellement des fonctions de test de synonymie et d'en expérimenter l'usage.

L'utilité des outils d'aide à la traduction reposant sur les mémoires de traduction est souvent limitée par la nature des segments que celles-ci mettent en correspondance, le plus souvent des phrases entières. Cet article examine le potentiel d'un type de système qui serait en mesure de récupérer la traduction de séquences de mots de longueur arbitraire.

Dans cet article, nous proposons une plate-forme multi-agents pour l'expérimentation et le traitement linguistique. Après une description du modèle d'agent APA, nous présentons l'état actuel de nos travaux: une implémentation en système multi-agents de l'analyse syntaxique selon le

paradigme des grammaires de dépendances en chunk. Nous montrons ensuite d'autres possibilités d'implémentation selon d'autres paradigmes syntaxiques mais aussi au delà de la simple syntaxe.

Cet article concerne la caractérisation et la représentation de la structure interne des énumérations. Pour ce faire, nous utilisons deux modèles de texte : d'une part la Théorie des Structures Rhétoriques (RST) qui fournit un cadre d'interprétation pour la structure discursive des textes et d'autre part le modèle de représentation de l'architecture textuelle qui est principalement dédié à l'étude et à la représentation des structures visuelles des textes. Après une brève présentation des modèles, nous nous concentrons sur l'étude de l'objet "énumérations". Nous exhibons et commentons trois exemples d'énumérations spécifiques que nous appelons des énumérations non-parallèles. Nous analysons la structure de ces énumérations et proposons un principe de composition des modèles de référence pour représenter ces énumérations. Enfin, nous présentons une classification des énumérations s'appuyant sur les caractéristiques de ces modèles.

L'ambiguïté syntaxique constitue un problème particulièrement délicat à résoudre pour les analyseurs morpho-syntaxiques des logiciels d'aide à la traduction, en particulier dans le cas des longs groupes nominaux typiques des langues de spécialité. En utilisant un corpus bilingue d'articles médicaux anglais traduits vers le français, nous examinons divers moyens de résoudre l'ambiguïté du rattachement de l'adjectif à l'un des deux noms qui le suivent dans les tournures anglaises de forme adjectif-nom-nom.

Cet article décrit un système d'extraction d'information sur les interactions entre gènes à partir de grandes bases de données textuelles. Le système est fondé sur une analyse au moyen de transducteurs à nombre fini d'états. L'article montre comment une partie des ressources (verbes d'interaction) peut être acquise de manière semi-automatique. Une évaluation détaillée du système est fournie.

Cet article présente un système pour l'identification automatique des expressions temporelles dans des textes français. La procédure d'identification repose sur une stratégie d'exploration contextuelle qui met en oeuvre deux techniques complémentaires: recherche des patrons (expressions régulières) et chart parsing qui est déclenché en fonction des patrons repérés.

Le cadre de cette étude concerne les systèmes de dialogue via le téléphone entre un serveur de données et un utilisateur. Nous nous intéresserons au cas de dialogues non contraints où l'utilisateur à toute liberté pour formuler ses requêtes. Généralement, le module de Reconnaissance Automatique de la Parole (RAP) de tels serveurs utilise un seul Modèle de Langage (ML) de type bi-gramme ou tri-gramme pour modéliser l'ensemble des interventions possibles de l'utilisateur. Ces ML sont appris sur des corpus de phrases retranscrites à partir de sessions entre le serveur et plusieurs utilisateurs. Nous proposons dans cette étude une méthode de segmentation de corpus d'apprentissage de dialogue utilisant une stratégie mixte basée à la fois sur des connaissances explicites mais aussi sur l'optimisation d'un critère statistique. Nous montrons qu'un gain en terme de perplexité et de taux d'erreurs/mot peut être constaté en utilisant un ensemble de sous modèles de langage issus de la segmentation plutôt qu'un modèle unique appris sur l'ensemble du corpus.

Cette contribution présente les ressources linguistiques informatisées du laboratoire ATILF (Analyses et Traitements Informatiques du Lexique Français) disponibles sur la toile et sert de support aux démonstrations prévues dans le cadre de TALN 2001. L'ATILF est la nouvelle U1\/[R créée en association entre le CNRS et l'Université Nancy 2 qui, depuis le 2 janvier 2001, a succédé à la composante nancéienne de l'INaLF. Ces importantes ressources sur la langue française regroupent un ensemble de plus de 3500 textes réunis dans Frantext et divers dictionnaires, lexiques et autres bases de données. Ces ressources exploitent les fonctionnalités du logiciel Stella, qui correspond à un véritable moteur de recherche dédié aux bases textuelles s'appuyant sur une

nouvelle théorie des objets textuels. La politique du laboratoire consiste à ouvrir très largement ses ressources en particulier au monde de la recherche et de l'enseignement.

Les résumés constitués de phrases extraites d'un texte contiennent souvent des mots inutiles, il est possible de les éliminer ou d'en réduire le nombre. Par un étude comparative des phrases du documents et des phrases correspondantes dans le résumé, cet article présente un inventaire partiel des unités qui sont souvent éliminées ou réduites en nombre. Les metadiscours, les textes entre parenthèses, les unités redondantes (emphases, répétitions), les appositions, modifieurs et relatives ne sont pas a place dans un résumé.

Nous présentons les avancées d'un projet dans un thème que nous qualifions de Cartographie de Textes qui permet à l'utilisateur novice d'explorer un nouveau domaine par navigation au sein d'un corpus homogène grâce à des cartes conceptuelles interactives. Une carte est composée de concepts pertinents relativement à la requête initiale et à son évolution, au sein du corpus; des relations extraites du corpus les lient aux mots de la requête. Des techniques d'apprentissage automatique sont combinées avec des heuristiques statistiques de Traitement Automatique des Langues pour la mise en évidence de collocations afin de construire les cartes.

Dans cette étude, nous proposons un modèle pour la résolution de la référence dans le cadre du dialogue homme machine. Partant de considérations psychologiques sur la nécessité d'un partage du système inférenciel pour permettre la communication, nous définissons un alisme basé sur des règles de production associées à des coûts cognitifs. Au travers d'exemples, nous montrons comment ce formalisme peut être utilisé comme cadre pour intégrer le traitement de différents phénomènes liés à la référence, et comment cette tégration peut conduire à des interfaces en langue naturelle plus efficaces.

Cet article discute de différentes approches pour faire le suivi automatique du courrier-électronique. Nous présentons tout d'abord les méthodes de traitement automatique de la langue (TAL) les plus utilisées pour cette tâche, puis un ensemble de critères influençant le choix d'une approche. Ces critères ont été développés grâce à une étude de cas sur un corpus fourni par Bell Canada Entreprises. Avec notre corpus, il est apparu que si aucune méthode n'est complètement satisfaisante par elle-même, une approche combinée semble beaucoup plus prometteuse.

Cet article porte sur l'identification de noms propres à partir de textes écrits. Les stratégies à base de règles développées pour des textes de type journalistique se révèlent généralement insuffisantes pour des corpus composés de textes ne répondant pas à des critères rédactionnels stricts. Après une brève revue des travaux effectués sur des corpus de textes de nature journalistique, nous présentons la problématique de l'analyse de textes variés en nous basant sur deux corpus composés de courriers électroniques et de transcriptions manuelles de conversations téléphoniques. Une fois les sources d'erreurs présentées, nous décrivons l'approche utilisée pour adapter un système d'extraction de noms propres développé pour des textes journalistiques à l'analyse de messages électroniques.

SEEK-JAVA est un système permettant Pidentification, l'interprétation et la représentation de connaissances à partir de textes. Il attribue une étiquette aux relations et identifie automatiquement les concepts arguments des relations. Les résultats, capitalisés dans une base de données, sont proposés, par le biais d'une interface, soit sous forme de graphes soit sous forme de tables. Ce système, intégré dans la plate-forme FilText, s'appuie sur la méthode d' exploration contextuelle.

Dans cet article, nous présentons un gestionnaire de dialogue pour un système de demande d'informations à reconnaissance vocale. Le gestionnaire de dialogue dispose de différentes sources de connaissance, des connaissances statiques et des connaissances dynamiques. Ces

connaissances sont gérées et utilisées par le gestionnaire de dialogue via des stratégies. Elles sont mises en oeuvre et organisées en fonction des objectifs concernant le système de dialogue et en fonction des choix ergonomiques que nous avons retenus. Le gestionnaire de dialogue utilise un modèle de dialogue fondé sur la détermination de phases et un modèle de la tâche dynamique. Il augmente les possibilités d'adaptation de la stratégie en fonction des historiques et de l'état du dialogue. Ce gestionnaire de dialogue, implémenté et évalué lors de la dernière campagne d'évaluation du projet LE-3 ARISE, a permi une amélioration du taux de succès de dialogue (de 53% à 85%).

Les modèles de langage stochastiques utilisés pour la reconnaissance de la parole continue, ainsi que dans certains systèmes de traitement automatique de la langue, favorisent pour la plupart l'interprétation d'un signal par les phrases les plus courtes possibles, celles-ci étant par construction bien souvent affectées des coûts les plus bas. Cet article expose un algorithme permettant de répondre à ce problème en remplaçant le coût habituel affecté par le modèle de langage par sa moyenne sur la longueur de la phrase considérée. Cet algorithme est très général et peut être adapté aisément à de nombreux modèles de langage, y compris sur des tâches d'analyse syntaxique.

Nous appliquons la désambiguïsation du sens des mot aux définitions d'un dictionnaire explicatif espagnol. Pour calculer le grand nombre de sens de mot en se basant sur le contexte (qui, dans notre cas, est la définition du dictionnaire), nous employons une modification de l'algorithme de Lesk. L'algorithme originel compare les mots pour savoir si ils appartiennent à un même lexème ou non; notre modification consiste en une comparaison floue employant un grand dictionnaire de synonyme et un système de morphologie dérivationnelle simple. L'application de la désambiguïsation aux définitions de dictionnaire (par contraste avec des textes habituels) permet quelques simplifications de l'algorithme (par exemple, nous ne nous soucions pas de la taille de la

fenêtre de contexte).

L'apport de connaissances linguistiques à la recherche d'information reste un sujet de débat. Nous examinons ici l'influence de connaissances morphologiques (flexion, dérivation) sur les résultats d'une tâche spécifique de recherche d'information dans un domaine spécialisé. Cette influence est étudiée à l'aide d'une liste de requêtes réelles recueillies sur un serveur opérationnel ne disposant pas de connaissances linguistiques. Nous observons que pour cette tâche, flexion et dérivation apportent un gain modéré mais réel.

On appelle grammaire de dépendance toute grammaire formelle qui manipule comme représentations syntaxiques des structures de dépendance. Le but de ce cours est de présenter à la fois les grammaires de dépendance (formalismes et algorithmes de synthése et d'analyse) et la théorie Sens-Texte, une théorie linguistique riche et pourtant méconnue, dans laquelle la dépendance joue un rôle crucial et qui sert de base théorique à plusieurs grammaires de dépendance.

Des groupes de recherche de plus en plus nombreux s'intéressent à l'étiquetage lexical ou la désambiguïsation du sens. La tendance actuelle est à l'exploitation de très grands corpus de textes qui, grâce à l'utilisation d'outils lexicographiques appropriés, peuvent fournir un ensemble de données initiales aux systèmes. A leur tour ces systèmes peuvent être utilisés pour extraire plus d'informations des corpus, qui peuvent ensuite être réinjectées dans les systèmes, dans un processus récursif. Dans cet article, nous présentons une méthodologie qui aborde la résolution de l'ambiguïté lexicale comme le résultat de l'interaction de divers indices repérables de manière semi-automatique au niveau syntaxique (valence), sémantique (collocations, classes d'objets) avec la mise en oeuvre de tests manuels.

Nous présentons dans cet article les débuts d'un travail visant à rechercher et à étudier systématiquement les critères de désambiguïsation sémantique automatique. Cette étude utilise un corpus français étiqueté sémantiquement dans le cadre du projet SyntSem. Le critère ici étudié est celui des co-occurrences. Nous présentons une série de résultats sur le pouvoir désambiguïsateur des co-occurrences en fonction de leur catégorie grammaticale et de leur éloignement du mot à désambiguïser.

Nous présentons un outil visant à assister les développeurs de ressources linguistiques en automatisant la fouille de corpus. Cet outil, est guidé par les principes de l'analyse distributionnelle sur corpus spécialisés, étendue grâce à des ressources lexicales génériques. Nous présentons une évaluation du gain de performances dû à l'intégration de notre outil à une application de filtrage d'information et nous élargissons le champ d'application de l'assistant aux études sur corpus menées à l'aide de cascades de transducteurs à états finis.

La nécessité de ressources lexicales normalisées et publiques est avérée dans le domaine du TAL. Cet article vise à montrer comment, sur la base d'une partie du lexique MULTEXT disponible sur le serveur ABU, il serait possible de construire une architecture permettant tout à la fois l'accès aux ressources avec des attentes différentes (lemmatiseur, parseur, extraction d'informations, prédiction, etc.) et la mise à jour par un groupe restreint de ces ressources. Cette mise à jour consistant en l'intégration et la modification, automatique ou manuelle, de données existantes. Pour ce faire, nous cherchons à prendre en compte à la fois les besoins et les données accessibles. Ce modèle est évalué conceptuellement dans un premier temps en fonction des systèmes utilisés dans notre équipe : un analyseur TAG, un constructeur de grammaires TAGs, un extracteur d'information.

Le but de cet article est d'expliciter certains des problèmes rencontrés lorsque l'on cherche à concevoir un système de reconnaissance de gestes de la Langue des Signes et de proposer des

solutions adaptées. Les trois aspects traités ici concernent la simultanéïté d'informations véhiculées par les gestes des mains, la synchronisation éventuelle entre les deux mains et le fait que différentes classes de signes peuvent se rencontrer dans une phrase en Langue des Signes.

Nous présentons dans cet article un cas particulier de description définie où la description reprend le rôle thématique d'un argument (implicite ou explicite) d'un événement mentionné dans le contexte linguistique. Nous commençons par montrer que les schémas d'annotation proposés (MATE) et utilisés (Poesio et Vieira 2000) ne permettent pas une caractérisation uniforme ni, partant, un repérage facile de ces reprises. Nous proposons une extension du schéma MATE qui pallie cette difficulté.

Cet article rapporte les résultats d'une étude quantitative des répétitions menée à partir d'un corpus de français parlé spontané d'un million de mots, étude réalisée dans le cadre de notre première année de thèse. L'étude linguistique pourra aider à l'amélioration des systèmes de reconnaissance de la parole et de l'étiquetage grammatical automatique de corpus oraux. Ces technologies impliquent la prise en compte et l'étude des répétitions de performance (en opposition aux répétitions de compétence, telles que nous nous sujet + complément) afin de pouvoir, par la suite, les « gommer » avant des traitements ultérieurs. Nos résultats montrent que les répétitions de performance concernent principalement les mots-outils et apparaissent à des frontières syntaxiques majeures.

La problématique de la normalisation de documents est introduite et illustrée par des exemples issus de notices pharmaceutiques. Un paradigme pour l'analyse du contenu des documents est proposé. Ce paradigme se base sur la spécification formelle de la sémantique des documents et utilise une notion de similarité floue entre les prédictions textuelles d'un générateur de texte et le texte du document à analyser. Une implémentation initiale du paradigme est présentée.

Nous proposons un modèle de conception d'agent conversationnel pour l'assistance d'interface. Notre but est d'obtenir un système d'interprétation de requêtes en contexte, générique pour l'indépendance vis-à-vis de la tâche, extensible pour sa capacité à intégrer des connaissances sur un nouveau domaine sans remettre en cause les connaissances antérieures et unifié dans le sens où tous les aspects du traitement de la langue naturelle, syntaxe, sémantique, ou pragmatique doivent s'exprimer dans un même formalisme. L'originalité de notre système est de permettre de représenter des connaissances d'interprétation de niveaux de granularité divers sous une même forme, réduisant la problématique de communication entre sources de connaissances qui existe dans les systèmes modulaires. Nous adoptons l'approche des micro-systèmes suivant laquelle l'interprétation de la langue se fait selon un processus non stratifié, et où absolument tous les niveaux peuvent interagir entre eux. Pour cela, nous introduisons et définissons un type d'entité que nous avons nommé observateur.

Nous présentons l'état de développement d'un outil d'extraction et de classification automatique de phrases pour la création de tests de langue. Cet outil de TAL est conçu pour, dans un premier temps, localiser et extraire de larges corpus en ligne du matériel textuel (phrases) possédant des propriétés linguistiques bien spécifiques. Il permet, dans un deuxième temps, de classifier automatiquement ces phrases-candidates d'après le type d'erreurs qu'elles sont en mesure de contenir. Le développement de cet outil s'inscrit dans un contexte d'optimalisation du processus de production d'items pour les tests d'évaluation. Pour répondre aux exigences croissantes de production, les industries de développement de tests de compétences doivent être capable de développer rapidement de grandes quantités de tests. De plus, pour des raisons de sécurité, les items doivent être continuellement remplacés, ce qui crée un besoin d'approvisionnement constant. Ces exigences de production et révision sont, pour ces organisations, coûteuses en temps et en personnel. Les bénéfices à retirer du développement et de l'implantation d'un outil capable

d'automatiser la majeure partie du processus de production de ces items sont par conséquents considérables.

Nous présentons dans cet article un logiciel d'étude permettant la création, la gestion et la manipulation de corpus de textes. Ce logiciel appelé MemLabor se veut un outil ouvert et open-source adaptable à toutes les opérations possibles que l'on peut effectuer sur ce type de matériau. Dans une première partie, nous présenterons les principes généraux de l'outil. Dans une seconde, nous en proposerons une utilisation dans le cadre d'une acquisition supervisée de classes sémantiques.

Les systèmes actuels de filtrage de l'information sont basés d'une façon directe ou indirecte sur les techniques traditionnelles de recherche d'information (Malone, Kenneth, 1987), (Kilander, Takkinen, 1996). Notre approche consiste à séparer le processus de classification du filtrage proprement dit. Il s'agit d'effectuer un traitement reposant sur une compréhension primitive du message permettant d'effectuer des opérations de classement. Cet article décrit une solution pour classer des messages en se basant sur les propriétés linguistiques véhiculées par ces messages. Les propriétés linguistiques sont modélisées par un réseau de neurone. A l'aide d'un module d'apprentissage, le réseau est amélioré progressivement au fur et à mesure de son utilisation. Nous présentons à la fin les résultats d'une expérience d'évaluation.

Qu'il s'adresse à un Prix Nobel ou à un étudiant de première année Maurice Gross ne craignait jamais d'être trop élémentaire. C'était à chaque fois comme si, entreprenant d'écrire un livre de mathématiques il ne pouvait rien démontrer avant d'avoir reconstruit les données les plus primitives du calcul et du raisonnement qui l'accompagne. Et il arrivait souvent que ceux qui l'écoutaient ou le lisaient pour la première fois, manquant par leur impatience le détail qui faisait que ses évidences n'avaient rien d'évident, s'imaginent qu'il les prenait pour des imbéciles. Parce qu'il avait l'expression

littéraire et philosophique, la langue du style, la forme de l'émotion, dans les tripes ? il pouvait citer sans discontinuer des poètes français ou anglais du XVIe siècle à nos jours et discuter longuement des formulations exactes d'un René Descartes ou d'un Charles Sanders Pierce, deux de ses deux philosophes préférés - il n'a jamais fait recette auprès des littéraires, des psycho-socios, des sémio-machins, des politiques et des pouvoirs académiques chez qui le raccourci, la connotation, le clin d'oeil, dont tout le monde a oublié sur quelles complicités exactes ils se fondent, tiennent lieu de découverte quand ce n'est pas de pensée. La complexité qui l'intéressait était d'une tout autre nature et autrement plus complexe. Elle avait pour horizon la phrase simple. Même pas l'énoncé, juste la phrase. Et simple c'est-à-dire constituée d'une seule proposition. Contrairement à ceux qui voyaient dans les processus de récursivité propositionnelle ? relatives notamment -- une source de complexité et de créativité, il y voyait un mécanisme très banal 1. La vraie complexité, celle qu'aucune machine construite à ce jour ne contrôle vraiment, il l'a exposée avec une simplicité désarmante en un peu moins de deux pages au début de Méthodes en syntaxe (1975: 17-19) dans le chapitre intitulé La créativité du langage. Elle porte sur les combinaisons possibles ou impossibles au sein d'une structure de neuf constituants formant une phrase simple. Mais ces possibilités "limitées à 1050 cas" et qui peuvent donc "être considérées comme intuitivement infinies" sans qu'il soit nécessaire "de faire appel à des mécanismes infinis pour rendre compte de leur richesse" ne sont qu'un horizon virtuel.

En recherche documentaire, on représente souvent les documents textuels par des vecteurs lexicaux de grande dimension qui sont redondants et coûteux. Il est utile de réduire la dimension des ces représentations pour des raisons à la fois techniques et sémantiques. Cependant les techniques classiques d'analyse factorielle comme l'ACP ne permettent pas de traiter des vecteurs de très grande dimension. Nous avons alors utilisé une méthode adaptative neuronale (GHA) qui s'est révélée efficace pour calculer un nombre réduit de nouvelles dimensions représentatives des données. L'approche nous a permis de classer un corpus réel de pages Web avec de bons

résultats.

Cette dernière décennie a été le témoin d'importantes avancées dans le domaine de la traduction statistique (TS). Aucune évaluation fine n'a cependant été proposée pour mesurer l'adéquation de l'approche statistique dans un contexte applicatif réel. Dans cette étude, nous étudions le comportement d'un engin de traduction probabiliste lorsqu'il traduit un texte de nature très éloignée de celle du corpus utilisé lors de l'entraînement. Nous quantifions en particulier la baisse de performance du système et développons l'idée que l'intégration de ressources terminologiques dans le processus est une solution naturelle et salutaire à la traduction. Nous décrivons cette intégration et évaluons son potentiel.

Certaines ressources textuelles ou terminologiques sont écrites sans signes diacritiques, ce qui freine leur utilisation pour le traitement automatique des langues. Dans un domaine spécialisé comme la médecine, il est fréquent que les mots rencontrés ne se trouvent pas dans les lexiques électroniques disponibles. Se pose alors la question de l'accentuation de mots inconnus : c'est le sujet de ce travail. Nous proposons deux méthodes d'accentuation de mots inconnus fondées sur un apprentissage par observation des contextes d'occurrence des lettres à accentuer dans un ensemble de mots d'entraînement, l'une adaptée de l'étiquetage morpho-syntaxique, l'autre adaptée d'une méthode d'apprentissage de règles morphologiques. Nous présentons des résultats expérimentaux pour la lettre e sur un thesaurus biomédical en français : le MeSH. Ces méthodes obtiennent une précision de 86 à 96 % (+-4 %) pour un rappel allant de 72 à 86 %.

Nous présentons une méthode d'analyse descendante et calculatoire. La démarche d'analyse est descendante du document à la proposition, en passant par la phrase. Le prototype présenté prend en entrée des documents en anglais, français, italien, espagnol, ou allemand. Il segmente les phrases en propositions, et calcule les relations sujet-verbe dans les propositions. Il est calculatoire,

car il exécute un petit nombre d'opérations sur les données. Il utilise très peu de ressources (environ 200 mots et locutions par langue), et le traitement de la phrase fait environ 60 Ko de Perl, ressources lexicales comprises. La méthode présentée se situe dans le cadre d'une recherche plus générale du Groupe Syntaxe et Ingénierie Mult-ilingue du GREYC sur l'exploration de solutions minimales et multi-lingues, ajustées à une tâche donnée, exploitant peu de propriétés linguistiques profondes, la généricité allant de pair avec l'efficacité.

Nous présentons un module mettant en oeuvre une méthode d'analyse distributionnelle dite "étendue". L'analyseur syntaxique de corpus SYNTEX effectue l'analyse en dépendance de chacune des phrases du corpus, puis construit un réseau de mots et syntagmes, dans lequel chaque syntagme est relié à sa tête et à ses expansions. A partir de ce réseau, le module d'analyse distributionnelle UPERY construit pour chaque terme du réseau l'ensemble de ses contextes syntaxiques. Les termes et les contextes syntaxiques peuvent être simples ou complexes. Le module rapproche ensuite les termes, ainsi que les contextes syntaxiques, sur la base de mesures de proximité distributionnelle. L'ensemble de ces résultats est utilisé comme aide à la construction d'ontologie à partir de corpus spécialisés.

Cet article survole les fonctionnalités offertes par le système DyALog pour construire des analyseurs syntaxiques tabulaires. Offrant la richesse d'un environnement de programmation en logique, DyALog facilite l'écriture de grammaires, couvre plusieurs formalismes et permet le paramétrage de stratégies d'analyse.

Les grammaires hors-contexte stochastiques sont exploitées par des algorithmes particulièrement efficaces dans des tâches de reconnaissance de la parole et d'analyse syntaxique. Cet article propose une autre probabilisation de ces grammaires, dont les propriétés mathématiques semblent intuitivement plus adaptées à ces tâches que celles des SCFG (Stochastique CFG), sans nécessiter

d'algorithme d'analyse spécifique. L'utilisation de ce modèle en analyse sur du texte provenant du corpus Susanne peut réduire de 33% le nombre d'analyses erronées, en comparaison avec une SCFG entraînée dans les mêmes conditions.

Nous présentons une méthode automatique d'extraction d'information à partir d'un corpus mono-domaine de mauvaise qualité, sur lequel il est impossible d'appliquer les méthodes classiques de traitement de la langue naturelle. Cette approche se fonde sur la construction d'une ontologie semi-formelle (modélisant les informations contenues dans le corpus et les relations entre elles). Notre méthode se déroule en trois phases : 1) la normalisation du corpus, 2) la construction de l'ontologie, et 3) sa formalisation sous la forme d'une grammaire. L'extraction d'information à proprement parler exploite un étiquetage utilisant les règles définies par la grammaire. Nous illustrons notre démarche d'une application sur un corpus bancaire.

Les modèles statistiques du langage ont pour but de donner une représentation statistique de la langue mais souffrent de nombreuses imperfections. Des travaux récents ont montré que ces modèles peuvent être améliorés s'ils peuvent bénéficier de la connaissance du thème traité, afin de s'y adapter. Le thème du document est alors obtenu par un mécanisme d'identification thématique, mais les thèmes ainsi traités sont souvent de granularité différente, c'est pourquoi il nous semble opportun qu'ils soient organisés dans une hiérarchie. Cette structuration des thèmes implique la mise en place de techniques spécifiques d'identification thématique. Cet article propose un modèle statistique à base d'uni-grammes pour identifier automatiquement le thème d'un document parmi une arborescence prédéfinie de thèmes possibles. Nous présentons également un critère qui permet au modèle de donner un degré de fiabilité à la décision prise. L'ensemble des expérimentations a été réalisé sur des données extraites du groupe 'fr' des forums de discussion.

Dans le cadre de recherches sur le sens en traitement automatique du langage, nous nous

concentrons sur la représentation de l'aspect thématique des segments textuels à l'aide de vecteurs conceptuels. Les vecteurs conceptuels sont automatiquement appris à partir de définitions issues de dictionnaires à usage humain (Schwab, 2001). Un noyau de termes manuellement indexés est nécessaire pour l'amorçage de cette analyse. Lorsque l'item défini s'y prête, ces définitions sont complétées par des termes en relation avec lui. Ces relations sont des fonctions lexicales (Mel'cuk and al, 95) comme l'hyponymie, l'hyperonymie, la synonymie ou l'antonymie. Cet article propose d'améliorer la fonction d'antonymie naïve exposée dans (Schwab, 2001) et (Schwab and al, 2002) grâce à ces informations. La fonction s'auto-modifie, par révision de listes, en fonction des relations d'antonymie avérées entre deux items. Nous exposons la méthode utilisée, quelques résultats puis nous concluons sur les perspectives ouvertes.

L'intégration de co-occurrences dans les modèles de représentation vectorielle de documents s'est avérée une source d'amélioration de la pertinence des mesures de similarités textuelles calculées dans le cadre de ces modèles (Rajman et al., 2000; Besançon, 2001). Dans cette optique, la définition des contextes pris en compte pour les co-occurrences est cruciale, par son influence sur les performances des modèles à base de co-occurrences. Dans cet article, nous proposons d'étudier deux méthodes de filtrage des co-occurrences fondées sur l'utilisation d'informations syntaxiques supplémentaires. Nous présentons également une évaluation de ces méthodes dans le cadre de la tâche de la recherche documentaire.

L'adaptation des modèles de langage dans les systèmes de reconnaissance de la parole est un des enjeux importants de ces dernières années. Elle permet de poursuivre la reconnaissance en utilisant le modèle de langage adéquat : celui correspondant au thème identifié. Dans cet article nous proposons une méthode originale de détection de thème fondée sur des vocabulaires caractéristiques de thèmes et sur la similarité entre mots et thèmes. Cette méthode dépasse la méthode classique (TF-IDF) de 14%, ce qui représente un gain important en terme d'identification.

Nous montrons également l'intérêt de choisir un vocabulaire adéquat. Notre méthode de détermination des vocabulaires atteint des performances 3 fois supérieures à celles obtenues avec des vocabulaires construits sur la fréquence des mots.

Nous exposons dans cet article une méthode réalisant de façon intégrée deux tâches de l'analyse thématique : la segmentation et la détection de liens thématiques. Cette méthode exploite conjointement la récurrence des mots dans les textes et les liens issus d'un réseau de collocations afin de compenser les faiblesses respectives des deux approches. Nous présentons son évaluation concernant la segmentation sur un corpus en français et un corpus en anglais et nous proposons une mesure d'évaluation spécifiquement adaptée à ce type de systèmes.

Le système de compréhension présenté dans cet article propose une approche logique et lexicalisée associant syntaxe et sémantique pour une analyse non sélective et hors-cadres sémantiques prédéterminés. L'analyse se déroule suivant deux grandes étapes ; un chunking est suivi d'une mise en relation des chunks qui aboutit à la construction de la représentation sémantique finale : formule logique ou graphe conceptuel. Nous montrons comment le formalisme a dû évoluer pour accroître l'importance de la syntaxe et améliorer la généricité des règles. Malgré l'utilisation d'une connaissance pragmatico-sémantique liée à l'application, la spécificité du système est circonscrite au choix des mots du lexique et à la définition de cette connaissance. Les résultats d'une campagne d'évaluation ont mis en évidence une bonne tolérance aux inattendus et aux phénomènes complexes, prouvant ainsi la validité de l'approche.

Cet article explore l'utilisation de ressources lexicales et textuelles ainsi que d'outils issus du TAL dans le domaine de l'apprentissage des langues assisté par ordinateur (ALAO). Il aborde le problème de la génération automatique ou semi-automatique d'exercices contextuels de vocabulaire à partir d'un corpus de textes et de données lexicales au moyen d'un étiqueteur et d'un parseur.

Sont étudiées les caractéristiques et les limites de ces exercices.

Cet article montre que pour une application telle qu'un système de question ? réponse, une analyse par mots clés de la question est insuffisante et qu'une analyse plus détaillée passant par une analyse syntaxique permet de fournir des caractéristiques permettant une meilleure recherche de la réponse.

Nous présentons dans cet article un cadre d'explication des relations entre les différents composants de l'analyse linguistique (prosodie, syntaxe, sémantique, etc.). Nous proposons un principe spécifiant un équilibre pour un objet linguistique donné entre ces différents composants sous la forme d'un poids (précisant l'aspect marqué de l'objet décrit) défini pour chacun d'entre eux et d'un seuil (correspondant à la somme de ces poids) à atteindre. Une telle approche permet d'expliquer certains phénomènes de variabilité : le choix d'une "tournure" à l'intérieur d'un des composants peut varier à condition que son poids n'empêche pas d'atteindre le seuil spécifié. Ce type d'information, outre son intérêt purement linguistique, constitue le premier élément de réponse pour l'introduction de la variabilité dans des applications comme les systèmes de génération ou de synthèse de la parole.

Dans cette étude, menée dans le cadre de la réalisation d'un analyseur syntaxique de corpus spécialisés, nous nous intéressons à la question des arguments et circonstants et à leur repérage automatique en corpus. Nous proposons une mesure simple pour distinguer automatiquement, au sein des groupes prépositionnels rattachés au verbe, des types de compléments différents. Nous réalisons cette distinction sur corpus, en mettant en oeuvre une stratégie endogène, et en utilisant deux mesures de productivité : la productivité du recteur verbal vis à vis de la préposition évalue le degré de cohésion entre le verbe et son groupe prépositionnel (GP), tandis que la productivité du régi vis à vis de la préposition permet d'évaluer le degré de cohésion interne du GP. Cet article

présente ces deux mesures, commente les données obtenues, et détermine dans quelle mesure cette partition recouvre la distinction traditionnelle entre arguments et circonstants.

La polysémie et la synonymie sont deux aspects fondamentaux de la langue. Nous présentons ici une évaluation de l'importance de ces deux phénomènes à l'aide de statistiques basées sur le lexique Word-Net et sur le SemCor. Ainsi, on a un taux de polysémie théorique de 5 sens par mot dans le SemCor. Mais si on regarde les occurrences réelles, moins de 50 % des sens possibles sont utilisés. De même, s'il y a, en moyenne, 2,7 mots possibles pour désigner un concept qui apparaît dans le corpus, plus de la moitié d'entre eux ne sont jamais utilisés. Ces résultats relativisent l'utilité de telles ressources sémantiques pour le traitement de la langue.

Cet article propose une méthode de codage automatique de traits lexicaux sémantiques en français. Cette approche exploite les relations fixées par l'instruction sémantique d'un opérateur de construction morphologique entre la base et le mot construit. En cela, la réflexion s'inspire des travaux de Marc Light (Light 1996) tout en exploitant le fonctionnement d'un système d'analyse morphologique existant : l'analyseur DériF. A ce jour, l'analyse de 12 types morphologiques conduit à l'étiquetage d'environ 10 % d'un lexique composé de 99000 lemmes. L'article s'achève par la description de deux techniques utilisées pour valider les traits sémantiques.

L'article présente Webaffix, un outil d'acquisition de couples de lexèmes morphologiquement apparentés à partir du Web. La méthode utilisé est inductive et indépendante des langues particulières. Webaffix (1) utilise un moteur de recherche pour collecter des formes candidates qui contiennent un suffixe graphémique donné, (2) prédit les bases potentielles de ces candidats et (3) recherche sur le Web des co-occurrences des candidats et de leurs bases prédites. L'outil a été utilisé pour enrichir Verbaction, un lexique de liens entre verbes et noms d'action ou d'événement correspondants. L'article inclut une évaluation des liens morphologiques acquis.

Cet article vise à évaluer deux approches différentes pour la constitution de classes sémantiques. Une approche endogène (acquisition à partir d'un corpus) est contrastée avec une approche exogène (à travers un réseau sémantique riche). L'article présente une évaluation fine de ces deux techniques.

Cet article présente une étude des conflits engendrés par la reconnaissance des entités nommées (EN) pour le français, ainsi que quelques indices pour les résoudre. Cette reconnaissance est réalisée par le système Nemesis, dont les spécifications ont été élaborées conséquemment à une étude en corpus. Nemesis se base sur des règles de grammaire, exploite des lexiques spécialisés et comporte un module d'apprentissage. Les performances atteintes par Nemesis, sur les anthroponymes et les toponymes, sont de 90% pour le rappel et 95% pour la précision.

La coédition d'un texte en langue naturelle et de sa représentation dans une forme interlingue semble le moyen le meilleur et le plus simple de partager la révision du texte vers plusieurs langues. Pour diverses raisons, les graphes UNL sont les meilleurs candidats dans ce contexte. Nous développons un prototype où, dans le scénario avec partage le plus simple, des utilisateurs "naïfs" interagissent directement avec le texte dans leur langue (L0), et indirectement avec le graphe associé pour corriger les erreurs. Le graphe modifié est ensuite envoyé au déconvertisseur UNL-L0 et le résultat est affiché. S'il est satisfaisant, les erreurs étaient probablement dues au graphe et non au déconvertisseur, et le graphe est envoyé aux déconvertisseurs vers d'autres langues. Les versions dans certaines autres langues connues de l'utilisateur peuvent être affichées, de sorte que le partage de l'amélioration soit visible et encourageant. Comme les nouvelles versions sont ajoutées dans le document multi-lingue original avec des balises et des attributs appropriés, rien n'est jamais perdu, et le travail coopératif sur un même document est rendu possible. Du côté interne, des liaisons sont établies entre des éléments du texte et du graphe en utilisant des

ressources largement disponibles comme un dictionnaire L0-anglais, ou mieux L0-UNL, un analyseur morpho-syntaxique de L0, et une transformation canonique de graphe UNL à arbre. On peut établir une "meilleure" correspondance entre "l'arbre-UNL+L0" et la "structure MS-L0", une treille, en utilisant le dictionnaire et en cherchant à aligner l'arbre et une trajectoire avec aussi peu que possible de croisements de liaisons. Un but central de cette recherche est de fusionner les approches de la TA par pivot, de la TA interactive, et de la génération multi-lingue de texte.

Nous présentons dans cet article le système de traduction français-anglais MSR-MT développé à Microsoft dans le groupe de recherche sur le traitement du language (NLP). Ce système est basé sur des analyseurs sophistiqués qui produisent des formes logiques, dans la langue source et la langue cible. Ces formes logiques sont alignées pour produire la base de données du transfert, qui contient les correspondances entre langue source et langue cible, utilisées lors de la traduction. Nous présentons différents stages du développement de notre système, commencé en novembre 2000. Nous montrons que les performances d'octobre 2001 de notre système sont meilleures que celles du système commercial Systran, pour le domaine technique, et décrivons le travail linguistique qui nous a permis d'arriver à cette performance. Nous présentons enfin les résultats préliminaires sur un corpus plus général, les débats parlementaires du corpus du Hansard. Quoique nos résultats ne soient pas aussi concluants que pour le domaine technique, nous sommes convaincues que la résolution des problèmes d'analyse que nous avons identifiés nous permettra d'améliorer notre performance.

Nous présentons un nouveau formalisme linguistique, les Grammaires d'Interaction, dont les objets syntaxiques de base sont des descriptions d'arbres, c'est-à-dire des formules logiques spécifiant partiellement des arbres syntaxiques. Dans ce contexte, l'analyse syntaxique se traduit par la construction de modèles de descriptions sous la forme d'arbres syntaxiques complètement spécifiés. L'opération de composition syntaxique qui permet cette construction pas à pas est

contrôlée par un système de traits polarisés agissant comme des charges électrostatiques.

Le système de question-réponse QALC utilise les documents sélectionnés par un moteur de recherche pour la question posée, les sépare en phrases afin de comparer chaque phrase avec la question, puis localise la réponse soit en détectant l'entité nommée recherchée, soit en appliquant des patrons syntaxiques d'extraction de la réponse, sortes de schémas figés de réponse pour un type donné de question. Les patrons d'extraction que nous avons définis se fondent sur la notion de focus, qui est l'élément important de la question, celui qui devra se trouver dans la phrase réponse. Dans cet article, nous décrirons comment nous déterminons le focus dans la question, puis comment nous l'utilisons dans l'appariement question-phrase et pour la localisation de la réponse dans les phrases les plus pertinentes retenues.

Cet article présente deux corpus francophones de dialogue oral (OTG et ECOLE_MASSY) mis librement à la disposition de la communauté scientifique. Ces deux corpus constituent la première livraison du projet Parole Publique initié par le laboratoire VALORIA. Ce projet vise la constitution d'une collection de corpus de dialogue oral enrichis par annotation morpho-syntaxique. Ces corpus de dialogue finalisé sont essentiellement destinés à une utilisation en communication homme-machine.

Notre objectif est de repérer l'existence de régularités pour prévoir l'attachement du relatif et de son référent introduit par un SN de la forme dét. N1 de (dét.) N2 en vue d'un traitement automatique. Pour évaluer les préférences, nous avons entrepris une analyse sur corpus. L'examen des occurrences examinées laisse entrevoir des variations en fonction de paramètres d'ordre fonctionnel, syntaxique et sémantiques : déterminants, traits sémantiques, saillance, relation établie par l'utilisation de de, position grammaticale du SN qui introduit le référent dans le discours?

Après avoir présenté le modèle computationnel de l'interprétation de métaphores proposé par Kintsch (2000), nous rapportons une étude préliminaire qui évalue son efficacité dans le traitement de métaphores littéraires et la possibilité de l'employer pour leur identification.

Un bon nombre des applications de traitement automatique des langues qui ont pour domaine les langues de spécialité sont des outils d'extraction terminologique. Elles se concentrent donc naturellement sur l'identification des groupes nominaux et des groupes prépositionnels ou prémodificateurs qui leur sont associés. En nous fondant sur un corpus composé d'articles de recherche médicale de langue anglaise, nous proposons un modèle d'extraction phraséologique semi-automatisée. Afin de distinguer, dans le cas des expressions de patron syntaxique <Adjectif ? Nom>, les termes de la langue médicale des simples collocations, nous nous sommes livré au repérage des adjectifs entrant en co-occurrence avec les adverbes. Cette méthode, qui permet l'élimination de la plupart des adjectifs relationnels, s'avère efficace en termes de précision. L'amélioration de son rappel nécessite toutefois l'utilisation de corpus de grande taille ayant subi un étiquetage morpho-syntaxique préalable.

Initialement développée au sein de l'université Paris VII depuis maintenant près de quinze ans, la grammaire FTAG, une implémentation du modèle des grammaires d'arbres adjoints pour le français, a connu ses dernières années, diverses évolutions majeures. (Candito, 1996) a ainsi réalisé l'intégration d'un modèle de représentation compact et hiérarchique d'informations redondantes que peut contenir une grammaire, au sein d'un système déjà existant. Ce modèle, que nous appelons MétaGammaire (MG) nous a permis, en pratique, de générer semi-automatiquement des arbres élémentaires, et par là même, d'augmenter de façon considérable les différents phénomènes syntaxiques couverts par notre grammaire. Un soin tout particulier a donc été apporté pour traiter les prédicats verbaux, en laissant cependant (partiellement) de côté le prédicat adjectival. Nous présentons donc ici une nouvelle implémentation de ce prédicat dans le cadre d'une extension de la

grammaire FTAG existante.

Le but de ce papier est de caractériser (au moins partiellement) les Grammaires à Substitution d'Arbres Polynômiales (pSTSG), instances particulières de STSG pour lesquelles la recherche de l'analyse la plus probable peut être effectuée en un temps polynômial. Nous donnons tout d'abord diverses conditions suffisantes, utilisables en pratique, qui garantissent qu'une STSG est polynômiale. Une telle condition suffisante, fondée sur la notion de "tête de syntagme", est ensuite présentée et évaluée.

Le domaine des "InterfacesUtilisateur Intelligentes" a vu ces dernières années la réalisation de systèmes complexes mettant en oeuvre une interaction multimodale dans laquelle les différentes techniques de communication (textes, gestes, parole, sélection graphique) sont coordonnées en entrée et/ou en sortie. Nous nous intéressons ici aux systèmes qui prennent en entrée des expressions multimodales et en produisent une reformulation en une expression unimodale sémantiquement équivalente. Nous proposons une modélisation du processus de traduction d'expressions multimodales en expressions unimodales, et nous décrivons la mise en oeuvre d'un processus de ce type dans un logiciel d'aide à l'apprentissage du langage.

Dans cet article, nous aborderons la notion d'attentes, vue du côté du locuteur, afin d'améliorer la modélisation du dialogue. Nous présenterons notre définition des attentes ainsi que notre notation, fondée sur une approche pragmatique du dialogue. Nous comparerons deux approches, l'une (uniquement stochastique) fondée sur la prédiction d'actes de parole, l'autre mettant en jeu les attentes du locuteur et leur gestion.

Nous présentons une approche de découpage thématique que nous utiliserons pour faciliter l'extraction d'information à partir de conversations téléphoniques transcrites. Nous expérimentons

avec un modèle de Markov caché utilisant des informations de différents niveaux linguistiques, des marques d'extra-grammaticalités et les entités nommées comme source additionnelle d'information. Nous comparons le modèle obtenu avec notre modèle de base utilisant uniquement les marques linguistiques et les extra-grammaticalités. Les résultats montrent l'efficacité de l'approche utilisant les entités nommées.

Partant du principe que certaines phrases peuvent réaliser plusieurs actes de langage, i.e., dans une interface sémantique?pragmatique, plusieurs constituants de discours séparés, nous proposons, dans le cadre de la SDRT, un algorithme de construction de représentations sémantiques qui prend en compte tous les aspects discursifs dès que possible et de façon compositionnelle.

Dans cet article, nous présentons un système de Compréhension Automatique de la Parole dont l'un des objectifs est de permettre un traitement fiable et robuste des inattendus structuraux du français parlé (hésitations, répétitions et corrections). L'analyse d'un énoncé s'effectue en deux étapes : une première étape générique d'analyse syntaxique de surface suivie d'une seconde étape d'analyse sémantico-pragmatique, dépendante du domaine d'application et reposant sur un formalisme lexicalisé : les grammaires de liens. Les résultats de l'évaluation de ce système lors de la campagne d'évaluation du Groupe de Travail Compréhension Robuste du GDR I3 du CNRS nous permettent de discuter de l'intérêt et des limitations de l'approche adoptée.

Nous proposons de présenter quelques-unes des ressources linguistiques informatisées que le laboratoire ATILF propose sur la toile et leurs diversités d'exploitation potentielle. Ces importantes ressources sur la langue française regroupent un ensemble de divers dictionnaires et lexiques, et de bases de données dont les plus importants sont le TLFi (Trésor de la Langue Française informatisé) et Frantext (plus de 3500 textes, dont la plupart catégorisés). Elles exploitent, pour la plupart, les

fonctionnalités du logiciel Stella, qui correspond à un véritable moteur de recherche dédié aux bases textuelles s'appuyant sur une nouvelle théorie des objets textuels. Tous les spécialistes de traitement automatique de la langue ainsi que tous les linguistes, syntacticiens aussi bien que sémanticiens, stylisticiens et autres peuvent exploiter avec bonheur les possibilités offertes par Stella sur le TLFi et autres ressources offertes par l'ATILF. Ces recherches peuvent s'articuler autour des axes suivants : études en vue de repérer des co-occurrences et collocations, extraction de sous-lexiques, études morphologiques, études de syntaxe locale, études de sémantique, études de stylistique, etc. Nous proposons de démystifier le maniement des requêtes sur le TLFi, FRANTEXT et nos autres ressources à l'aide du logiciel Stella, et d'expliquer et de montrer comment interroger au mieux ces ressources et utiliser l'hyper-navigation mise en place entre ces ressources pour en tirer les meilleurs bénéfices.

Ce tutoriel est une introduction à la modélisation lexicographique des liens lexicaux au moyen des fonctions lexicales de la théorie Sens-Texte. Il s'agit donc d'examiner un sous-ensemble des tâches effectuées en lexicographie formelle basée sur la lexicologie explicative et combinatoire. Plutôt que de viser l'introduction de toutes les fonctions lexicales identifiées par la théorie Sens- Texte, je vais m'attacher à introduire la notion de fonction lexicale de façon méthodique, en présentant d'abord les notions linguistiques plus générales sur lesquelles elle s'appuie (lexie, prédicat, actant, dérivation sémantique, collocation, etc.). Ce document vise essentiellement à récapituler les définitions des notions linguistiques qui vont être vues dans le tutoriel de façon pratique, par le biais d'exercices à caractère lexicographique.

Nous présenterons tout d'abord la philosophie « Agents » en général, afin d'en montrer les avantages pour le domaine du TALN, qui se caractérise par une hétérogénéité avérée des systèmes existants (multiplicité des langages de programmation), ainsi qu'une forte demande en ressources (mémoire notamment). Nous ferons ensuite une présentation des principales

plate-formes orientées agents, puis nous examinerons de plus près la plate-forme développée au Standford Research Institute (SRI) : OAA (licence libre). Nous clôturerons le tutoriel sur des exemples commentés d'applications industrielles utilisant OAA, permettant de donner toutes les clés nécessaires au développement d'applications distribuées (intra/internet), multi-agents et multiplates-formes (plusieurs langages de programmation/systèmes d'exploitation).

De nombreux systèmes de Traitement Automatique des Langues (TAL) utilisent une architecture séquentielle basée sur la transmission, à la fin de chaque phase d'analyse, des résultats trouvés à la phase d'analyse suivante. Ces types de systèmes séquentiels posent plusieurs problèmes (i.e. explosion combinatoire des solutions, lourdeur d'analyse, etc.). Pour remédier à ces problèmes, plusieurs solutions de remplacement ont vu le jour, nous pouvons citer par exemple, l'utilisation des approches multi-agent que nous avons adopté pour faire l'analyse syntaxique de textes Arabes, et que nous présentons dans cet article.

L'objectif de cet article est de présenter nos travaux sur l'analyse d'un énoncé vers une structure de dépendance. Cette structure décrit les relations entre mots, des relations syntaxiques mais également des relations sémantiques de surface de l'énoncé de départ dans un certain contexte. L'idée est de créer une plate-forme d'analyse capable d'intégrer des analyseurs linguistiques existants (syntaxiques ou de dépendance) et de fusionner leurs résultats dans le but d'obtenir une analyse de dépendance pour des énoncés quelconques.

Dans cet article, nous présentons une méthode qui vise à donner à un utilisateur la possibilité de parcourir rapidement un ensemble de documents par le biais d'un profil utilisateur. Un profil est un ensemble de termes structuré en sous-ensembles thématiquement homogènes. L'analyse des documents se fonde pour sa part sur l'extraction des passages les plus étroitement en relation avec ce profil. Cette analyse permet en particulier d'étendre le vocabulaire définissant un profil en

fonction du document traité en sélectionnant les termes de ce dernier les plus étroitement liés aux termes du profil. Cette capacité ouvre ainsi la voie à une plus grande finesse du filtrage en permettant la sélection d'extraits de documents ayant un lien plus ténu avec les profils mais davantage susceptibles d'apporter des informations nouvelles et donc intéressantes. La production du résumé résulte de l'appariement entre les segments délimités lors de l'analyse des documents et les thèmes du profil.

Dans cet article, nous montrons que la cohérence d'un discours dépend de la relation entre la structure communicative des phrases et la structure du discours. Du point de vue de la synthèse, la visée communicative contrôle la structure du discours, et la structure du discours contraint le choix des structures communicatives phrastiques : nous proposons de reproduire ce processus dans un système de génération de textes. Nous montrons de quelle manière la structure communicative intervient lors de la phase de structuration de document pour permettre la génération de discours cohérents et répondant à des visées communicatives particulières.

Dans cette étude, nous nous intéressons à l'apport de ressources exogènes dans un analyseur syntaxique de corpus basé sur des procédures d'apprentissage endogène. Nous menons une expérience en corpus sur un cas d'ambiguïté catégorielle du français (forme de en position postverbale, article ou préposition). Après avoir présenté et évalué la stratégie endogène, nous en analysons les limites. Nous discutons ensuite la perspective d'une approche mixte combinant des informations acquises de manière endogène à des informations exogènes (données de sous-catégorisation verbale sur la préposition de). Nous montrons alors comment un apport maximal de ressources exogènes améliore les performances de l'analyseur (+8%, +15% sur les deux corpus évalués). Nous présentons les premiers résultats d'une approche mixte avant de conclure sur les orientations futures du travail.

Dans l'analyse sémantique de textes, un des obstacles au TAL est la polysémie des unités linguistiques. Par exemple, le sens du verbe jouer peut varier en fonction du contexte : Il joue de la trompette (pratiquer) ; Il joue avec son fils (s'amuser). Une des approches pour traiter ces ambiguïtés de sens, est le modèle de la construction dynamique du sens proposé par B. Victorri et C. Fuchs (1996). Dans ce modèle, on associe à chaque unité polysémique un espace sémantique, et le sens de l'unité dans un énoncé donné est le résultat d'une interaction dynamique avec les autres unités présentes dans l'énoncé. Nous voulons montrer ici que les constructions verbales sont des éléments du co-texte qui contribuent, au même titre que le co-texte lexical, au processus dynamique de construction du sens du verbe. L'objectif est alors de montrer que les constructions verbales sont porteuses de sens intrinsèque (Goldberg, 1995) et qu'elles permettent dans notre modèle de contraindre automatiquement le sens d'un verbe.

Dans cet article, nous montrons l'insuffisance du pouvoir d'expression des approches par prédicats pour la résolution de la référence en extension dans un cadre générique de dialogue homme-machine. Cette insuffisance oblige pour l'instant les concepteurs de tels systèmes de dialogue à concevoir des heuristiques ad hoc impossibles à intégrer dans un cadre de description unifié. Nous montrons que la résolution des expressions référentielles nécessite la prise en compte du contexte même pour les termes portant sur des caractéristiques intrinsèques aux éléments. Nous proposons alors un formalisme pour représenter la sémantique des extracteurs référentiels intrinsèques. Ce formalisme repose sur trois fonctions, la première permet de calculer le rapport de similarité de deux éléments en fonction d'une certaine dimension et dans un certain contexte, les deux autres permettent de partitionner un domaine de référence trié par l'utilisation de la première fonction.

Dans le cadre de la représentation du sens en TALN, nous développons actuellement un système d'analyse des aspects thématiques des textes et de désambiguïsation lexicale basée sur les

vecteurs conceptuels. Ces vecteurs visent à représenter un ensemble d'idées associées à tout segment textuel. À partir de ce modèle, nous avons posé des hypothèses sur la construction des vecteurs. Dans cet article, nous montrons comment ces hypothèses, ainsi que des considérations techniques comme la possibilité de distribuer les tâches à effectuer ou la modularité, nous ont amenées à adopter une architecture multi-agents. Chaque agent possède un certain nombre de compétences, une mémoire qui lui est propre et peut interragir avec son environnement (les autres agents). Pour finir, nous présentons les agents déjà implémentés et un exemple de leur collaboration.

Nous présentons dans cet article une réflexion en vue de la modélisation d'une partie du patrimoine descriptif du français finalement peu utilisé en TALN. Pour ce faire, nous utilisons le concept de langage "pivot" qui permet d'articuler la description et la présentation formalisée.

Cet article expose la recherche effectuée dans le cadre de mon doctorat visant à élaborer un étiquetage morphologique de l'anglais et à désambiguïser automatiquement les ambiguïtés dues à la morphologie dans le cadre du projet LABELGRAM [9]. Nous montrons qu'il est très pertinent et efficace de travailler conjointement sur l'étiquetage et la désambiguïsation. Nous décrivons de manière précise notre contribution au système qui a consisté à mettre en place la partie anglaise. Pour ce faire, nous avons établi un dictionnaire en intention, nous avons évalué quantitativement le phénomène d'ambiguïté morphologique et établi la validité de la méthode de désambiguïsation par règles contextuelles pour l'anglais.

Cette étude présente un modèle pour le traitement de la morphologie du finnois. Ce modèle est fondé sur des transducteurs à nombre fini d'états. L'approche utilise une façon originale d'organiser les données et de générer dynamiquement une structure sémantique à partir d'une analyse morphologique. L'approche est linguistiquement validée par une étude des suffixes de dérivation

verbale en finnois.

Nous présentons dans cet article un outil graphique de développement de grammaire, basé sur le formalisme des Grammaires de Propriétés. Nous y exprimons les raisons pour lesquelles l'association d'une représentation complète et ergonomique, et d'un modèle formel flexible et homogène fournit un avantage considérable pour l'intégration des informations issues de la linguistique descriptive.

Dans le cadre de la recherche en sémantique lexicale, nous utilisons le modèle des vecteurs conceptuels pour représenter les sens de termes. La base vectorielle est construite à partir de définitions provenant de diverses sources lexicales, ce qui permet statistiquement de tempérer les diverses incohérences locales. Pour désigner le sens obtenu après un regroupement des définitions, nous utilisons un identificateur qui entraîne certaines contraintes. En particulier, un "cluster" de définition est désigné par une référence vers différentes définitions de la multisource. D'autre part, le contrôle de la qualité d'une classification ou désambiguisation de sens impose de faire référence en permanence au lexique source. Nous proposons donc de nommer un sens à l'aide d'un autre terme du lexique. L'annotation est un outil léger et efficace qui est essentiellement une association d'idées que l'on peut extraire de toute base de connaissance linguistique. Les annotations obtenues peuvent finalement constituer une nouvelle source d'apprentissage pour la base de vecteurs conceptuels.

Un des problèmes rencontrés lors de l'analyse de textes en chinois est qu'il n'existe pas de séparateur entré lés mots dans cette langue. Le mot étant une unité linguistique fondamentale en traitement automatique dé la langue, il est nécessaire d'identifier les mots dans un texte chinois afin que des analysés de plus haut niveau puissent être réalisées. Le but de cet article est dé présenter un système d'idéntification dés mots basé sur un algorithme utilisant des triplets dé catégories

grammaticales ét dés fréquences de mots. Cé système comprend deux dictionnaires : l'un dédié aux mots ét à léurs fréquences, l'autré aux triplets dés catégories correspondantes. Les tests qui ont été effectués révèlent que 98,5% dés phrases sont découpées correctement. Certaines erreurs sont dués à la taillé limitée du dictionnaire utilisé. Une réflexion sur la création de nouvelles catégories ét dés études proposant des règles grammaticales sont en cours de réalisation afin d'aug1nénter la performance du système.

Cet article présente une architecture générique de système de dialogue oral homme-machine. Premièrement, nous abordons quelques problèmes soulevés par la généricité des systèmes de dialogue homme-machine. Nous décrivons ensuite dans ce cadre quelques systèmes récents et typiques. Nous présentons finalement une architecture générique pour concevoir/construire des systèmes de dialogue oral homme-machine.

La plupart des systèmes de filtrage du courrier électronique existants enregistrent des lacunes ou faiblesses sur l'efficacité du filtrage. Certains systèmes sont basés seulement sur le traitement de la partie structurée (un ensemble de règles sur l'entête du message), et d'autres sont basés sur un balayage superficiel de la partie texte du message (occurrence d'un ensemble de mots clés décrivant les intérêts de l'utilisateur). Cet article propose une double amélioration de ces systèmes. D'une part, nous proposons un ensemble de critères automatisables et susceptibles d'influer sur le processus de filtrage. Ces critères sont des indices qui portent généralement sur la structure et le contenu des messages. D'autre part, nous utilisons une méthode d'apprentissage automatique permettant au système d'apprendre à partir de données et de s'adapter à la nature des mails dans le temps. Dans cet article, nous nous intéressons à un type de messages bien particulier, qui continue à polluer nos boîtes emails de façon croissante : les messages indésirables, appelés spam. Nous présentons à la fin les résultats d'une expérience d'évaluation.

Dans cet article, nous discutons de lapplication au langage parlé des techniques danalyse syntaxique robuste développées pour lécrit. Nous présentons deux systèmes de compréhension de parole spontané en situation de dialogue homme-machine finalisé, dont les performances montrent la pertinence de ces méthodes pour atteindre une compréhension fine et robuste des énoncés oraux.

Nous présentons dans cet article une étude sur les critères de désambiguïsation sémantique automatique basés sur les co-occurrences. Lalgorithme de désambiguïsation utilisé est du type liste de décision, il sélectionne une co-occurrence unique supposée véhiculer l'information la plus fiable dans le contexte ciblé. Cette étude porte sur 60 vocables répartis, de manière égale, en trois classes grammaticales (nom, adjectif et verbe) avec une granularité fine au niveau des sens. Nous commentons les résultats obtenus par chacun des critères évalués de manière indépendante et nous nous intéressons aux particularités qui différencient les trois classes grammaticales étudiées. Cette étude sappuie sur un corpus français étiqueté sémantiquement dans le cadre du projet SyntSem.

En reconnaissance de la parole, un des moyens d'améliorer les performances des systèmes est de passer par l'adaptation des modèles de langage. Une étape cruciale de ce processus consiste à détecter le thème du document traité et à adapter ensuite le modèle de langage. Dans cet article, nous proposons une nouvelle approche de création des vocabulaires utilisés pour la détection de thème. Cette dernière est fondée sur le développement de vocabulaires spécifiques et caractéristiques des différents thèmes. Nous montrons que cette approche permet non seulement d'améliorer les performances des méthodes, mais exploite également des vocabulaires de taille réduite. De plus, elle permet d'améliorer de façon très significative les performances de méthodes de détection lorsqu'elles sont combinées.

L'hypothèse soutenue dans cet article est que l'analyse de contenu, quand elle est réalisée par un analyseur syntaxique robuste avec calcul sémantique dans un modèle adéquat, est un outil de classification tout aussi performant que les méthodes statistiques. Pour étudier les possibilités de cette hypothèse en matière de classification, à l'aide de l'analyseur du Français, SYGMART, nous avons réalisé un projet en grandeur réelle avec une société qui propose des sélections d'articles en revue de presse. Cet article présente non seulement les résultats de cette étude (sur 4843 articles finalement sélectionnés), mais aussi cherche à montrer que l'analyse de contenu automatisée, quand elle est possible, est un moyen fiable de produire une catégorisation issue du sens (quand il est calculable), et pas simplement créée à partir d'une reconnaissance de "similarités" de surface.

Dans le modèle du Lexique génératif (Pustejovsky, 1995), certaines propriétés sémantiques des noms sont exprimées à l'aide de verbes. Les couples nom-verbe ainsi formés présentent un intérêt applicatif notamment en recherche d'information. Leur acquisition sur corpus constitue donc un enjeu, mais la découverte des patrons qui les définissent en contexte est également importante pour la compréhension même du modèle du Lexique génératif. Cet article présente une technique entièrement automatique permettant de répondre à ce double besoin d'extraction sur corpus de couples et de patrons morpho-syntaxiques et sémantiques. Elle combine pour ce faire deux approches d'acquisition? l'approche statistique et l'approche symbolique? en conservant les avantages propres à chacune d'entre elles : robustesse et automatisation des méthodes statistiques, qualité et expressivité des résultats des techniques symboliques.

Dans cet article, nous présentons un ensemble doutils de conception et dexploitation pour des grammaires darbres adjoints lexicalisés. Ces outils sappuient sur une représentation XML des ressources (lexique et grammaire). Dans notre représentation, à chaque arbre de la grammaire est associé un hypertag décrivant les phénomènes linguistiques quil recouvre. De ce fait, la liaison avec le lexique se trouve plus compactée et devient plus aisée à maintenir. Enfin, un analyseur permet

de valider les grammaires et les lexiques ainsi conçus aussi bien de façon interactive que différée sur des corpus.

Dans la tâche de désambiguïsation sémantique, la détermination de la taille optimale de fenêtre de contexte à utiliser, a fait l'objet de plusieurs études. Dans cet article, nous proposons une approche à deux niveaux pour répondre à cette problématique de manière automatique. Trois systèmes concurrents à base d'arbres de classification sémantique sont, dans un premier temps, utilisés pour déterminer les trois sens les plus vraisemblables d'un mot. Ensuite, un système décisionnel tranche entre ces sens au regard d'un contexte plus étendu. Les améliorations constatées lors d'expériences menées sur les données de SENSEVAL-1 et vérifiées sur les données SENSEVAL-2 sont significatives.

Cet article concerne les phrases complexes avec deux conjonctions de subordination. Nous montrerons que de telles phrases peuvent s'interpréter de quatre façons différentes. Il s'agit donc de formes fortement ambiguës pour lesquelles il est opportun d'avoir recours à des représentations sémantiques sous-spécifiées, et c'est ce que nous proposerons.

La fiabilité des réponses qu'il propose, ou un moyen de l'estimer, est le meilleur atout d'un système de question-réponse. A cette fin, nous avons choisi d'effectuer des recherches dans des ensembles de documents différents et de privilégier des résultats qui sont trouvés dans ces différentes sources. Ainsi, le système QALC travaille à la fois sur une collection finie d'articles de journaux et sur le Web.

Dans cet article, nous présentons une méthodologie d'apprentissage faiblement supervisé pour l'extraction automatique de paraphrases à partir du Web. À partir d'un seule exemple de paire (prédicat, arguments), un corpus est progressivement accumulé par sondage duWeb. Les phases de sondage alternent avec des phases de filtrage, durant lesquelles les paraphrases les moins

plausibles sont éliminées à l'aide d'une procédure de clustering non supervisée. Ce mécanisme d'apprentissage s'appuie sur un système de Questions-Réponses existant et les paraphrases apprises seront utilisées pour en améliorer le rappel. Nous nous concentrons ici sur le mécanisme d'apprentissage de ce système et en présentons les premiers résultats.

Dans cet article, nous développons les modules syntaxique et topologique du modèle Sens- Texte et nous montrons lutilité de la topologie comme représentation intermédiaire entre les représentations syntaxique et phonologique. Le modèle est implémenté dans un générateur et nous présentons la grammaire du grec moderne dans cette approche.

On présente une nouvelle variante de grammaire contextuelle structurée, qui produit des arbres de dépendance. Le nouveau modèle génératif, appelé grammaire contextuelle de dépendance, améliore la puissance générative forte et faible des grammaires contextuelles, tout en étant un candidat potentiel pour la description mathématique des modèles syntactiques de dépendance.

Cet article présente la normalisation de la sortie d'un analyseur robuste de l'anglais. Nous montrons quels sont les enrichissements que nous avons effectués afin de pouvoir obtenir à la sortie de notre analyseur des relations syntaxiques plus générales que celles que nous offrent habituellement les analyseurs robustes existants. Pour cela nous utilisons non seulement des propriétés syntaxiques, mais nous faisons appel aussi à de l'information de morphologie dérivationnelle. Cette tâche de normalisation est menée à bien grâce à notre analyseur XIP qui intègre tous les traitements allant du texte brut tout venant au texte normalisé. Nous pensons que cette normalisation nous permettra de mener avec plus de succès des tâches d'extraction d'information ou de détection de similarité entre documents.

Cet article présente tout dabord une analyse linguistique des cadres organisationnels et son

implémentation informatique. Puis à partir de ce travail, une modélisation généralisable à lensemble des cadres de discours est proposée. Enfin, nous discutons du concept dindicateur proposé dans le cadre théorique de lexploration contextuelle.

La compréhension automatique de la parole peut être considérée comme un problème d'association entre deux langages différents. En entrée, la requête exprimée en langage naturel et en sortie, juste avant l'étape d'interprétation, la même requête exprimée en terme de concepts. Un concept représente un sens bien déterminé. Il est défini par un ensemble de mots partageant les mêmes propriétés sémantiques. Dans cet article, nous proposons une méthode à base de réseau bayésien pour l'extraction automatique des concepts ainsi que trois approches différentes pour la représentation vectorielle des mots. Ces représentations aident un réseau bayésien à regrouper les mots, construisant ainsi la liste adéquate des concepts à partir d'un corpus d'apprentissage. Nous conclurons cet article par la description d'une étape de post-traitement au cours de laquelle, nous étiquetons nos requêtes et nous générons les commandes SQL appropriées validant ainsi, notre approche de compréhension.

Cet article présente une grammaire d'unification dans laquelle les morphèmes grammaticaux sont traités similairement aux morphèmes lexicaux!: les deux types de morphèmes sont traités comme des signes à part entière et sont décris par des structures élémentaires qui peuvent s'unifier directement les unes aux autres (ce qui en fait une grammaire de dépendance). Nous illustrerons notre propos par un fragment de l'interface sémantique-syntaxe du français pour le verbe et l'adjectif!: voix, modes, temps, impersonnel et tough-movement.

Que ce soit pour la compréhension ou pour la génération d'expressions référentielles, la Théorie de la Pertinence propose un critère cognitif permettant de comparer les pertinences de plusieurs expressions dans un contexte linguistique. Nous voulons ici aller plus loin dans cette voie en

proposant une caractérisation précise de ce critère, ainsi que des pistes pour sa quantification. Nous étendons l'analyse à la communication multimodale, et nous montrons comment la perception visuelle, le langage et le geste ostensif interagissent dans la production d'effets contextuels. Nous nous attachons à décrire l'effort de traitement d'une expression multimodale à l'aide de traits. Nous montrons alors comment des comparaisons entre ces traits permettent d'exploiter efficacement le critère de pertinence en communication homme-machine. Nous soulevons quelques points faibles de notre proposition et nous en tirons des perspectives pour une formalisation de la pertinence.

En dépit des travaux réalisés cette dernière décennie dans le cadre général de la traduction probabiliste, nous sommes toujours bien loin du jour où un engin de traduction automatique (probabiliste ou pas) sera capable de répondre pleinement aux besoins d'un traducteur professionnel. Dans une étude récente (Langlais, 2002), nous avons montré comment un engin de traduction probabiliste pouvait bénéficier de ressources terminologiques extérieures. Dans cette étude, nous montrons que les techniques de traduction probabiliste peuvent être utilisées pour extraire des informations sous-phrastiques d'une mémoire de traduction. Ces informations peuvent à leur tour s'avérer utiles à un engin de traduction probabiliste. Nous rapportons des résultats sur un corpus de test de taille importante en utilisant la mémoire de traduction d'un concordancier bilingue commercial.

Cet article propose une nouvelle classification des utilisations des démonstratifs, une mise en oeuvre de cette classification dans une analyse de corpus et présente les resultats obtenus au terme de cette analyse. La classification proposée est basée sur celles existant dans la littérature et étendues pour permettre la génération de groupes nominaux démonstratifs. L'analyse de corpus montre en particulier que la nature "reclassifiante" du démonstratif lui permet d'assumer deux fonctions (une fonction anaphorique et une fonction de support pour de l'information nouvelle) et qu'il existe des moyens variés de réaliser ces fonctions.

Dans cet article, nous proposons de montrer que la combinaison de plusieurs analyses syntaxiques permet d'extraire Panalyse la plus fiable pour une phrase donnée. De plus, chaque information syntaxique sera affectée d'un score de confiance déterminé selon le nombre d'analyseurs syntaxiques la confirmant. Nous verrons que cette approche implique l'étude des différents analyseurs syntaxiques existants ainsi que leur évaluation.

Les grammaires stochastiques standards utilisent des modèles probabilistes de nature générative, fondés sur des probabilités de récriture conditionnées par le symbole récrit. Les expériences montrent qu'elles tendent ainsi par nature à pénaliser les dérivations les plus longues pour une mˆeme entrée, ce qui n'est pas forcément un comportement souhaitable, ni en analyse syntaxique, ni en reconnaissance de la parole. Dans cet article, nous proposons une approche probabiliste non-générative du modèle STSG (grammaire stochastique à substitution d'arbres), selon laquelle les probabilités sont conditionnées par les feuilles des arbres syntaxiques plutˆot que par leur racine, et qui par nature fait appel à un apprentissage discriminant. Plusieurs expériences sur ce modèle sont présentées.

Dans le cadre du projet Papillon qui vise à la construction de bases lexicales multi-lingues par acceptions, nous avons défini des stratégies pour peupler un dictionnaire pivot de liens interlingues à partir d'une base vectorielle monolingue. Il peut y avoir un nombre important de sens par entrée et donc l'identification des acceptions correspondantes peut être erronée. Nous améliorons l'intégrité de la base d'acception grâce à des agents experts dans les fonctions lexicales comme la synonymie, l'antonymie, l'hypéronymie ou l'holonymie. Ces agents sont capable de calculer la pertinence d'une relation sémantique entre deux acceptions par les diverses informations lexicales récoltées et les vecteurs conceptuels. Si une certaine pertinence est au-dessus d'un seuil, ils créent un lien sémantique qui peut être utilisé par d'autres agents chargés par exemple de la

désambiguïsation ou du transfert lexical. Les agents vérifiant l'intégrité de la base cherchent les incohérences de la base et en avertissent les lexicographes le cas échéant.

Cette communication présente la version pour le français d'Amalgam, un système de réalisation automatique de phrases. Deux des modèles du système sont décrits en détail, et nous expliquons comment la performance des modèles peut être améliorée en combinant connaissances et intuition linguistiques et méthodes statistiques.

Nous présenterons dans cette communication les premiers travaux de modélisation informatique d'une grammaire de la langue créole martiniquaise, en nous inspirant des descriptions fonctionnelles de Damoiseau (1984) ainsi que du manuel de Pinalie & Bernabé (1999). Prenant appui sur des travaux antérieurs en génération de texte (Vaillant, 1997), nous utilisons un formalisme de grammaires d'unification, les grammaires d'adjonction d'arbres (TAG d'après l'acronyme anglais), ainsi qu'une modélisation de catégories lexicales fonctionnelles à base syntaxico-sémantique, pour mettre en oeuvre une grammaire du créole martiniquais utilisable dans une maquette de système de génération automatique. L'un des intérêts principaux de ce système pourrait être son utilisation comme logiciel outil pour l'aide à l'apprentissage du créole en tant que langue seconde.

Nous décrivons un algorithme, HyperLex, de détermination automatique des différents usages dun mot dans une base textuelle sans utilisation dun dictionnaire. Cet algorithme basé sur la détection des composantes de forte densité du graphe des co-occurrences de mots permet, contrairement aux méthodes précédemment proposées (vecteurs de mots), disoler des usages très peu fréquents. Il est associé à une technique de représentation graphique permettant à lutilisateur de naviguer de façon visuelle à travers le lexique et dexplorer les différentes thématiques correspondant aux usages discriminés.

Le filtrage de contenus illicites sur Internet est une problématique difficile qui est actuellement résolue par des approches à base de listes noires et de mots-clés. Les systèmes de classification textuelle par apprentissage automatique nécessitant peu d'interventions humaines, elles peuvent avantageusement remplacer ou compléter les méthodes précédentes pour faciliter les mises à jour. Ces techniques, traditionnellement utilisées avec des catégories définies par leur sujet (économie ou sport par exemple), sont fondées sur la présence ou l'absence de mots. Nous présentons une évaluation de ces techniques pour le filtrage de contenus racistes. Contrairement aux cas traditionnels, les documents ne doivent pas être catégorisés suivant leur sujet mais suivant le point de vue énoncé (raciste ou antiraciste). Nos résultats montrent que les classifieurs, essentiellement lexicaux, sont néanmoins bien adaptées : plus de 90% des documents sont correctement classés, voir même 99% si l'on accepte une classe de rejet (avec 20% d'exemples non classés).

Nous proposons une méthode pour apprendre des relations morphologiques dérivationnelles en corpus. Elle se fonde sur la co-occurrence en corpus de mots formellement proches et un filtrage complémentaire sur la forme des mots dérivés. Elle est mise en oeuvre et expérimentée sur un corpus médical. Les relations obtenues avant filtrage ont une précision moyenne de 75,6 % au 5000è rang (fenêtre de 150 mots). L'examen détaillé des dérivés adjectivaux d'un échantillon de 633 noms du champ de l'anatomie montre une bonne précision de 85?91 % et un rappel modéré de 32?34 %. Nous discutons ces résultats et proposons des pistes pour les compléter.

Le but des systèmes de question-réponse est de trouver des réponses exactes et factuelles à des questions exprimées en langue naturelle en recherchant dans une grande collection de documents. Notre recherche vise plutôt à générer des réponses complètes, sous forme de phrases, étant donnée la réponse exacte. La génération de telles phrases-réponses est une tâche importante car ces phrases peuvent être employées par un système de question-réponse pour améliorer la

recherche de réponses exactes ou bien, pour améliorer l'interface entre le système et l'utilisateur en fournissant des réponses plus naturelles. Suite à une étude de corpus de phrases réponses, nous avons développé un ensemble de patrons syntaxiques de réponses correspondant à chaque patron syntaxique de question.

La grammaire FTAG du français a vu ces dernières années ses données s'accroître très fortement. D'abord écrits manuellement, les arbres qui la composent, ont ensuite été générés semi-automatiquement grâce à une Métagrammaire, développée tout spécialement. Après la description des verbes en 1999, puis celle des adjectifs en 2001-2002, c'est maintenant au tour des verbes supports et des noms prédicatifs de venir enrichir les descriptions syntaxiques de la grammaire. Après un rappel linguistique et technique des notions de verbe support et de méta-grammaire, cet article présente les choix qui ont été entrepris en vue de la description de ces nouvelles données.

Cet article traite de l'acquisition automatique des grammaires de Lambek, utilisées pour la modélisation syntaxique des langues. Récemment, des algorithmes ont été proposés dans le modèle d'apprentissage de Gold, pour certaines classes de grammaires catégorielles. En revenche, les grammaires de Lambek rigides ou k-valuées ne sont pas apprenables à partir des chaînes. Nous nous intéressons ici au cas des grammaires de prégroupe. Nous montrons que la classe des grammaires de prégroupe n'est pas apprenable à partir des chaînes, même si on limite fortement l'ordre des types (ordre 1/2) ; notre preuve revient à construire un point limite pour cette classe.

Cet article concerne la structuration automatique de documents par des méthodes linguistiques. De telles procédures sont rendues nécessaires par les nouvelles tâches de recherche d'information intradocumentaires (systèmes de questions-réponses, navigation sélective dans des documents...). Nous développons une méthode exploitant la théorie de l'encadrement du discours de Charolles,

avec une application visée en recherche d'information dans les documents géographiques - d'où l'intérêt tout particulier porté aux cadres spatiaux et temporels. Nous décrivons une implémentation de la méthode de délimitation de ces cadres et son exploitation pour une tâche d'indexation intratextuelle croisant les critères spatiaux et temporels avec des critères thématiques.

Cet article fournit des éléments d'explication pour la description des relations entre les différents domaines de l'analyse linguistique. Il propose une architecture générale en vue d'une théorie formée de plusieurs niveaux : d'un côté les grammaires de chacun des domaines et de l'autre des relations spécifiant les interactions entre ces domaines. Dans cette approche, chacun des domaines est porteur d'une partie de l'information, celle-ci résultant également de l'interaction entre les domaines.

Cet article présente une application qui associe un certain nombre de valeurs sémantiques à des segments textuels en vue de proposer un traitement automatique de la temporalité dans les textes. Il s'agit d'automatiser une analyse sémantique de surface à l'aide de règles heuristiques d'exploration contextuelle et d'une base organisée de marqueurs linguistiques.

Nous présentons dans ses grandes lignes un modèle de structuration de documents pour la génération automatique de preuves mathématiques. Le modèle prend en entrée des sorties d'un prouveur automatique et vise à produire des textes dont le style s'approche le plus possible des démonstrations rédigées par des humains. Cela implique la mise au point d'une stratégie de planification de document capable de s'écarter de la structure purement logique de la preuve. La solution que nous proposons consiste à intégrer de manière simple des informations de type intentionnel afin d'enrichir la structure rhétorique finale du texte.

La reconnaissance de termes dans les textes intervient dans de nombreux domaines du Traitement Automatique des Langues Naturelles, qu'il s'agisse d'indexation automatique, de traduction, ou

d'extraction de connaissances. Nous présentons une méthodologie d'évaluation de Systèmes de Reconnaissance de Termes (SRT) qui vise à minimiser le temps d'expertise des spécialistes en faisant coopérer des SRT. La méthodologie est mise en oeuvre sur des textes en anglais dans le domaine de la chimie des métaux et à l'aide de deux SRT : FASTR et SYRETE. Le banc de test construit selon cette méthodologie a permis de valider les SRT et d'évaluer leurs performances en termes de rappel et de précision.

Les réseaux lexicaux de type Word-Net présentent une absence de relations de nature thématique, relations pourtant très utiles dans des tâches telles que le résumé automatique ou l'extraction d'information. Dans cet article, nous proposons une méthode visant à construire automatiquement à partir d'un large corpus un réseau lexical dont les relations sont préférentiellement thématiques. En l'absence d'utilisation de ressources de type dictionnaire, cette méthode se fonde sur un principe d'auto-amorçage : un réseau de collocations est d'abord construit à partir d'un corpus puis filtré sur la base des mots du corpus que le réseau initial a permis de sélectionner. Nous montrons au travers d'une évaluation portant sur la segmentation thématique que le réseau final, bien que de taille bien inférieure au réseau initial, permet d'obtenir les mêmes performances que celui-ci pour cette tâche.

Dans cet article, nous proposons une méthode non supervisée d'apprentissage qui permet d'améliorer la désambiguïsation du rattachement prépositionnel dans le cadre d'un analyseur robuste à base de règles pour le français. Les rattachements ambigus d'une première analyse sont transformés en requêtes sur leWeb dans le but de créer un grand corpus qui sera analysé et d'où seront extraites automatiquement des informations lexicales et statistiques sur les rattachements. Ces informations seront ensuite utilisées dans une deuxième analyse pour lever les ambiguïtés des rattachements. L'avantage d'une telle méthode est la prise en compte de co-occurrences syntaxiques et non pas des co-occurrences purement textuelles. En effet, les mesures statistiques (poids) sont associées à des mots apparaissant initialement dans une même relation de

dépendance, c'est-à-dire, des attachements produits par le parseur lors d'une première analyse.

Cet article présente un prototype de Question/Réponse (Q/R) impliquant un ensemble de bases de connaissances (BC) dont l'objectif est d'apporter un crédit supplémentaire aux réponses candidates trouvées. Ces BC et leur influence sur la stratégie dordonnancement mise en œuvre sont décrites dans le cadre de la participation du système à la campagne Q/R de TREC-2002.

Nous présentons dans ce travail un logiciel de mise au point de grammaires pour le traitement morpho-syntaxique de l'arabe et l'établissement de grammaires pour le filtrage et l'extraction d'information en arabe. Ce logiciel est fondé sur le principe des automates. L'analyse morpho-syntaxique de l'arabe est réalisé sans le recours au lexique.

Nous décrivons dans cet article l'utilisation darbres décisionnels pour lacquisition dinformations lexicales et lenrichissement de notre système de traitement automatique des langues naturelles (NLP). Notre approche diffère dautres projets dapprentissage automatique en ce quelle repose sur lexploitation dun système danalyse linguistique profonde. Après lintroduction de notre sujet nous présentons larchitecture de notre module dapprentissage lexical. Nous présentons ensuite une situation dapprentissage lexical effectué en utilisant des arbres décisionnels; nous apprenons quels verbes prennent un sujet humain en espagnol et en français.

Nous décrivons dans cet article l'implantation d'un système de rédaction contrôlée multi-lingue dans un environnement XML. Avec ce système, un auteur rédige interactivement un texte se conformant à des règles de bonne formation aux niveaux du contenu sémantique et de la réalisation linguistique décrites par un schéma XML. Nous discutons les avantages de cette approche ainsi que les difficultés rencontrées lors du développement de ce système. Nous concluons avec un exemple d'application à une classe de documents pharmaceutiques.

Dans ce travail, nous étudions lapport dun modèle de langage pour améliorer les performances des systèmes de reconnaissance de lécriture manuscrite en-ligne. Pour cela, nous avons exploré des modèles basés sur des approches statistiques construits par apprentissage sur des corpus écrits. Deux types de modèles ont été étudiés : les modèles n-grammes et ceux de type n-classes. En vue de lintégration dans un système de faible capacité (engin nomade), un modèle n-classe combinant critères syntaxiques et contextuels a été défini, il a permis dobtenir des résultats surpassant ceux donnés avec un modèle beaucoup plus lourd de type n-gramme. Les résultats présentés ici montrent quil est possible de prendre en compte les spécificités dun langage en vue de reconnaître lécriture manuscrite avec des modèles de taille tout à fait raisonnable.

L'objectif de cette contribution est de présenter l'intégration de la notion d'évaluation dans la méthodologie de prototypage rapide de modèles de dialogue développée et mise en oeuvre dans le cadre du projet InfoVox. L'idée centrale de cette méthodologie est de dériver un modèle de dialogue opérationnel directement à partir du modèle de la tâche à laquelle il est associé. L'intégration systématique de différents aspects de l'évaluation dans le processus de prototypage est alors utile afin d'identifier, dès la phase de conception, les qualités et défauts de l'interface. Toutes les conclusions présentées seront illustrées par des résultats concrets obtenus au cours d'expériences réalisées dans le cadre du projet InfoVox.

Cet article présente une méthode danalyse systématique et scientifique des documents constituant un dossier dinstruction. Lobjectif de cette approche est de pouvoir donner au juge dinstruction de nouveaux moyens pour évaluer la cohérence, les incohérences, la stabilité ou les variations dans les témoignages. Cela doit lui permettre de définir des pistes pour mener de nouvelles investigations. Nous décrivons les travaux que nous avons réalisés sur un dossier réel puis nous proposons une méthode danalyse des résultats.

Les mots arabes sont lexicalement beaucoup plus proches les uns des autres que les mots français et anglais. Cette proximité a pour effet un grand nombre de propositions à la correction d'une forme erronée arabe. Nous proposons dans cet article une méthode qui prend en considération le contexte de l'erreur pour éliminer certaines propositions données par le correcteur. Le contexte de l'erreur sera dans un premier temps les mots voisinant l'erreur et s'étendra jusqu'à l'ensemble des mots du texte contenant l'erreur. Ayant été testée sur un corpus textuel contenant des erreurs réelles, la méthode que nous proposons aura permis de réduire le nombre moyen de propositions d'environ 75% (de 16,8 à 3,98 propositions en moyenne).

Le nombre d'approches en traduction automatique s'est multiplié dans les dernières années. Il existe entre autres la traduction par règles, la traduction statistique et la traduction guidée par l'exemple. Dans cet article je decris les approches principales en traduction automatique. Je distingue les approches qui se basent sur des règles obtenues par l'inspection des approches qui se basent sur des exemples de traduction. La traduction guidée par l'exemple se caractérise par la phrase comme unité de traduction idéale. Une nouvelle traduction est génerée par analogie : seulement les parties qui changent par rapport à un ensemble de traductions connues sont adaptées, modifiées ou substituées. Je présente quelques techniques qui ont été utilisées pour ce faire. Je discuterai un système spécifique, EDGAR, plus en detail. Je démontrerai comment des textes traduits alignés peuvent être preparés en termes de compilation pour extraire des unités de traduction sous-phrastiques. Je présente des résultats en traduction Anglais -> Français produits avec le système EDGAR en les comparant avec ceux d'un système statistique.

Cet article constitue le support dun cours présenté lors de la conférence TALN 2003. Il défend la place du Traitement Automatique des Langues comme discipline clé pour le développement de ressources termino-ontologiques à partir de textes. Les contraintes et enjeux de ce processus sont

identifiés, en soulignant limportance de considérer cette tâche comme un processus supervisé par un analyste. Sont présentés un certain nombre doutils logiciels et méthodologiques venant de plusieurs disciplines comme le TAL et lingénierie des connaissances qui peuvent aider lanalyste dans sa tâche. Divers retours dexpérience sont présentés.

Dans cet article, nous présentons le système de Question Réponse QALC, et nous nous intéressons tout particulièrement à l'extraction de la réponse. Un appariement question-réponse fondé sur les relations syntaxiques a été développé, afin d'améliorer les performances du système. Un projet de génération de réponses à partir de plusieurs documents est également discuté.

Nous montrons dans cet article qu'un pré-étiquetage des usages des mots par un algorithme de désambiguïsation tel qu'HyperLex (Véronis, 2003, 2004) permet d'obtenir des relations lexicales (du type NOM-ADJECTIF, NOM de NOM, NOM-VERBE) beaucoup plus exploitables, parce qu'elles-mêmes catégorisées en fonction des usages. De plus, cette technique permet d'obtenir des relations pour des usages très peu fréquents, alors qu'une extraction indifférenciée « noie » ces relations au milieu de celles correspondant aux usages les plus fréquents. Nous avons conduit une évaluation sur un corpus de plusieurs milliers de pages Web comportant l'un des 10 mots-cibles très polysémiques choisis pour cette expérience, et nous montrons que la précision obtenue est très bonne, avec un rappel honorable, suffisant en tout cas pour de nombreuses applications. L'analyse des erreurs ouvre des perspectives d'améliorations pour la suite de notre travail de thèse.

Nous présentons ici le système d'indexation automatique actuellement en cours de développement dans l'équipe CISMeF afin d'aider les documentalistes lors de l'indexation de ressources de santé. Nous détaillons l'architecture du système pour l'extraction de mots clés MeSH, et présentons les résultats d'une première évaluation. La stratégie d'indexation choisie atteint une précision comparable à celle des systèmes existants. De plus, elle permet d'extraire des paires mot

clé/qualificatif, et non des termes isolés, ce qui constitue une indexation beaucoup plus fine. Les travaux en cours s'attachent à étendre la couverture des dictionnaires, et des tests à plus grande échelle sont envisagés afin de valider le système et d'évaluer sa valeur ajoutée dans le travail quotidien des documentalistes.

Nous présentons une méthode d'appariement de mots, à partir de corpus français/anglais alignés, qui s'appuie sur l'analyse syntaxique en dépendance des phrases. Tout d'abord, les mots sont appariés à un niveau global grâce au calcul des fréquences de co-occurrence dans des phrases alignées. Ces mots constituent les couples amorces qui servent de point de départ à la propagation des liens d'appariement à l'aide des différentes relations de dépendance identifiées par un analyseur syntaxique dans chacune des deux langues. Pour le moment, cette méthode dite d'appariement local traite majoritairement des cas de parallélisme, c'est-à-dire des cas où les relations syntaxiques sont identiques dans les deux langues et les mots appariés de même catégorie. Elle offre un taux de réussite de 95,4% toutes relations confondues.

Dans cet article, nous introduisons une approche de la représentation et de l'analyse des discours multimodaux, basée sur un traitement unimodulaire par contraintes. Le but de cet article est de présenter (i) un système de représentation des données et (ii) une méthode d'analyse, permettant une interaction simplifiée entre les différentes modalités de communication. L'avantage de cette méthode est qu'elle permet la prise en compte rigoureuse d'informations communicatives de natures diverses en un traitement unique, grâce à une représentation homogène des objets, de leurs relations, et de leur méthode d'analyse, selon le modèle des Grammaires de Propriétés.

Dans cet article, nous présentons une typologie des phénomènes qui posent problème pour l'annotation syntaxique de corpus oraux. Nous montrons également que ces phénomènes, même s'ils y sont d'une fréquence moindre, sont loin d'être absents à l'écrit (ils peuvent même être tout à

fait significatifs dans certains corpus : e-mails, chats, SMS?), et que leur prise en compte peut améliorer l'annotation et fournir un cadre intégré pour l'oral et l'écrit.

Nous présentons les automates lexicaux avec structure de traits, une extension du modèle des automates finis sur le mots dans lesquels les transitions sont étiquetées par des motifs qui sélectionnent un sous-ensemble des mots étiquetés en fonction de leurs traits positionnés. Nous montrons l'adéquation de ce modèle avec les ressources linguistiques dont nous disposons et nous exposons les grandes lignes de nos méthodes pour effectuer des opérations telles que la déterminisation, l'intersection ou la complémentation sur ces objets. Nous terminons en présentant une application concrète de ces méthodes pour la levée d'ambiguïtés lexicales par intersection d'automates à l'aide de contraintes locales.

Les recherches en sémantique lexicale s'appuient de plus en plus sur des ressources électroniques de grande taille (dictionnaires informatisés, corpus, ontologies) à partir desquelles on peut obtenir diverses relations sémantiques entre unités lexicales. Ces relations sont naturellement modélisées par des graphes. Bien qu'ils décrivent des phénomènes lexicaux très différents, ces graphes ont en commun des caractéristiques bien particulières. On dit qu'ils sont de type petit monde. Nous voulons mener une étude théorique mathématique et informatique de la structure de ces graphes pour le lexique. Il s'agit de les géométriser afin de faire apparaître l'organisation du lexique, qui est implicitement encodée dans leur structure. Les outils mis en place sont testés sur le graphe du dictionnaire électronique des synonymes (www.crisco.unicaen.fr). Ils constituent une extension du logiciel Visusyn développé par Ploux & Victorri (1998).

Cet article présente le projet de l'atelier logique ICHARATE dédié à l'étude des grammaires catégorielles multimodales. Cet atelier se présente sous la forme de bibliothèques pour l'assistant de preuves Coq.

Cet article s'intéresse aux définitions formalisées de la base de données BDéf et montre en quoi la structure formelle de ces définitions est à même d'offrir une représentation originale de la polysémie lexicale.

Dans cet article nous présentons un outil d'extraction d'information dédié à la veille qui répond à un certain nombre de requêtes formulées par l'utilisateur, en combinant la puissance des outils et les ressources informatiques à une analyse linguistique. Cette analyse linguistique permet le repérage des entités nommées (acteurs, lieux, temps,?) ainsi que la mise en relation des acteurs avec leur environnement dans l'espace et le temps au moyen d'indices déclencheurs, d'indices complémentaires et de règles qui les combinent, c'est le principe de l'Exploration Contextuelle. Les résultats capitalisés dans des fichiers XML, sont proposés par le biais d'une interface, soit sous forme de graphes soit sous forme de base d'informations.

Nous décrivons notre méthode de production automatique du résumé de textes juridiques. C'est une nouvelle application du résumé qui permet aux juristes de consulter rapidement les idées clés d'une décision juridique pour trouver les jurisprudences pertinentes à leurs besoins. Notre approche est basée sur l'exploitation de l'architecture des documents et les structures thématiques, afin de constituer automatiquement des fiches de résumé qui augmentent la cohérence et la lisibilité du résumé. Dans cet article nous détaillons les conceptions des différentes composantes du système, appelé LetSum et le résultat d'évaluation.

Pour la recherche documentaire il est souvent intéressant d'avoir une bonne mesure de confiance dans les réponses trouvées par le moteur de recherche. Une bonne estimation de pertinence peut permettre de faire un choix entre plusieurs réponses (venant éventuellement de différents systèmes), d'appliquer des méthodes d'enrichissement additionnelles selon les besoins, ou encore

de permettre à l'utilisateur de prendre des décisions (comme d'approfondir la recherche à travers un dialogue). Nous proposons une méthode permettant de faire une telle estimation, utilisant des connaissances extraites d'un ensemble de requêtes connues pour en déduire des prédictions sur d'autres requêtes posées au système de recherche documentaire.

Nous présentons une méthode multilingue de catégorisation en mot vide / mot plein à partir de corpus brut. Cette méthode fait appel à des propriétés très générales des langues ainsi qu'à des techniques issues de la communauté de la fouille de données.

Dans cet article, nous présentons un processus de génération sémantico-syntaxique conçu et mis en oeuvre dans la réalisation d'un prototype de traduction automatique basée sur le modèle à structure intermédiaire (ou structure pivot). Dans une première partie de l'article, nous présentons l'organisation des ressources lexicales et sémantiques multilingues, ainsi que les mécanismes permettant d'exploiter ces ressources pour produire une représentation conceptuelle du sens de la phrase source. Dans une seconde partie, nous présentons la première phase de génération à partir d'une structure pivot (génération Sémantico-Syntaxique) permettant la construction d'une structure syntaxique profonde de la phrase cible à produire. Les autres phases de génération ne seront pas abordées dans cet article.

L'analyse morphologique automatique du slovaque constitue la première étape d'un système d'analyse automatique du contenu des textes scientifiques et techniques slovaques. Un tel système pourrait être utilisé par des applications telles que l'indexation automatique des textes, la recherche automatique de la terminologie ou par un système de traduction. Une description des régularités de la langue par un ensemble de règles ainsi que l'utilisation de tous les éléments au niveau de la forme du mot qui rendent possible son interprétation permettent de réduire d'une manière considérable le Volume des dictionnaires. Notamment s'il s'agît d'une langue à flexion très riche,

comme le slovaque. La reconnaissance automatique des adjectifs durs et des adverbes réguliers constitue la partie la plus importante de nos travaux. Les résultats que nous obtenons lors de l'analyse morphologique confirment la faisabilité et la grande fiabilité d'une analyse morphologique basée sur la reconnaissance des formes et ceci pour toutes les catégories lexicales.

Afin de générer les formes fléchies des noms luxembourgeois dans le dictionnaire luxembourgeois, nous utilisons un code flexionnel. Ce code s'étant révélé trop contraignant pour traiter l'inflexion (alternance vocalique/Umlaut), nous présentons ici un moyen efficace pour coder ce phénomène. La pertinence de ce type de code est double. D'une part, il correspond mieux aux besoins du linguiste qui aimerait établir des classes flexionnelles naturelles sans trop de contraintes informatiques. D'autre part, il permet de réduire significativement le nombre de classes flexionnelles. Le dictionnaire électronique luxembourgeois dispose ainsi de deux codes qui peuvent se combiner entre eux pour mieux traiter les particularités morphologiques des mots luxembourgeois.

Par ses caractéristiques éminentes dans la présentation des données, Self-Organizing Map (SOM) est particulièrement convenable à l'organisation des cartes. SOM se comporte d'un ensemble des vecteurs prototypes pour représenter les données d'entrée, et fait une projection, en conservant la topologie, à partir des vecteurs prototypes de n-dimensions sur une carte de 2-dimensions. Cette carte deviendra une vision qui reflète la structure des classes des données. Nous notons un problème crucial pour SOM, c'est la méthode de vectorisation des données. Dans nos études, les données se présentent sous forme des textes. Bien que le modèle général du SOM soit déjà créé, il nous faut de nouvelles recherches pour traiter des langues spécifiques, comme le vietnamien, qui sont de nature assez différente de l'anglais. Donc, nous avons appliqué la conception du syntagme pour établir un algorithme qui est capable de résoudre ce problème.

L'objectif de notre travail est de dégager une représentation formelle compositionnelle de la

contribution sémantique de aussi lorsqu'il a une valeur additive. Plusieurs problèmes de compositionnalité, liés surtout à la diversité des arguments concernés par l'adverbe, vont se poser. Nous proposons une alternative compositionnelle à la représentation proposée initialement en l-DRT.

Notre article présente les composants nécessaires à la synthèse de la parole arabe. Nous nous attarderons sur la transcription graphème phonème, étape primordiale pour l'élaboration d'un système de synthèse d'une qualité acceptable. Nous présenterons ensuite quelques-unes des règles utilisées pour la réalisation de notre système de traitement phonétique. Ces règles sont, pour notre système, stockées dans une base de données et sont parcourues plusieurs fois lors de la transcription.

Les significations des expressions dans les langues naturelles sont souvent indéterminées (sous-spécifiées) et nécessitent d'être enrichies avant de devenir des propositions complètes. La sémantique générale des expressions linguistiques doit être complétée par les inférences pragmatiques, identifiées et captées d'une manière régulière et permettant ainsi un traitement opérationnel et même informatique. Cet article étudie l'indétermination de l'aspect imperfectif en russe et propose un cadre sémantique et pragmatique pour l'identification de ses différentes valeurs sémantiques à la base de règles.

Cet article présente une stratégie de construction semi-automatique d'une base lexicale interlingue par acception, à partir de ressources existantes, qui utilise en synergie des techniques existantes de désambiguïsation. Les apports et limitations de chaque technique sont présentés. Notre proposition est de pouvoir composer arbitrairement des techniques, en fonction des ressources disponibles, afin d'obtenir une base interlingue de la qualité souhaitée. Jeminie, un système adaptable qui met en oeuvre cette stratégie, est introduit dans cet article.

Il est souhaitable qu'une analyse syntaxique -en traitement automatique des langues naturelles soit réalisée avec plus ou moins de précision en fonction du contexte, c'est-à-dire que sa granularité soit réglable. Afin d'atteindre cet objectif, nous présentons ici des études préliminaires permettant d'appréhender les contextes technique et scientifique qui soulèvent ce problème. Nous établissons un cadre pour les développements à réaliser. Plusieurs types de granularité sont définis. Puis nous décrivons une technique basée sur la densité de satisfaction, développée dans ce cadre avec des algorithmes basés sur un formalisme de satisfaction de contraintes (celui des Grammaires de Propriétés) ayant l'avantage de permettre l'utilisation des mêmes ressources linguistiques avec un degré de précision réglable. Enfin, nous envisageons les développements ultérieurs pour une analyse syntaxique à granularité variable.

Nous présentons ici une méthode de réutilisation de systèmes de traduction automatique gratuits en ligne pour développer des applications multilingues et évaluer ces mêmes systèmes. Nous avons développé un outil de traitement et de traduction de documents hétérogènes (multilingues et multicodage). Cet outil permet d'identifier la langue et le codage du texte, de segmenter un texte hétérogène en zones homogènes, d'appeler un traducteur correspondant avec une paire de langue source et cible, et de récupérer les résultats traduits dans la langue souhaitée. Cet outil est utilisable dans plusieurs applications différentes comme la recherche multilingue, la traduction des courriers électroniques, la construction de sites web multilingues, etc.

L'accès au contenu des textes de génomique est aujourd'hui un enjeu important. Cela suppose au départ d'identifier les noms d'entités biologiques comme les gènes ou les protéines. Se pose alors la question de la variation de ces noms. Cette question revêt une importance particulière en génomique où les noms de gènes sont soumis à de nombreuses variations, notamment la synonymie. A partir d'une étude de corpus montrant que la synonymie est une relation stable et

linguistiquement marquée, cet article propose une modélisation de la synonymie et une méthode d'extraction spécifiquement adaptée à cette relation. Au vu de nos premières expériences, cette méthode semble plus prometteuse que les approches génériques utilisées pour l'extraction de cette relation.

S'inscrivant dans les domaines du TAL, de la linguistique sur corpus et de l'informatique documentaire, l'étude présentée ici opère plus précisément dans la perspective d'une analyse macro-sémantique de la structuration discursive. Plus spécifiquement, nous proposons une analyse sémantique des structures rhétoriques du discours. Après avoir envisagé certaines voies ouvertes en la matière, nous définissons notre approche, et présentons les expérimentations conduites, dans le cadre du projet GeoSem, sur les structures énumératives dans le domaine géographique.

Nous proposons une évaluation de différentes méthodes et outils de segmentation thématique de textes. Nous présentons les outils de segmentation linéaire et non supervisée DotPlotting, Segmenter, C99, TextTiling, ainsi qu'une manière de les adapter et de les tester sur des documents français. Les résultats des tests montrent des différences en performance notables selon les sujets abordés dans les documents, et selon que le nombre de segments à trouver est fixé au préalable par l'utilisateur. Ces travaux font partie du projet Technolangue AGILE-OURAL 1.

Nous présentons les résultats d'expérimentations visant à introduire des ressources lexico-syntaxiques génériques dans un analyseur syntaxique de corpus à base endogène (SYNTEX) pour la résolution d'ambiguïtés de rattachement prépositionnel. Les données de sous-catégorisation verbale sont élaborées à partir du lexique-grammaire et d'une acquisition en corpus (journal Le Monde). Nous présentons la stratégie endogène de désambiguïsation, avant d'y intégrer les ressources construites. Ces stratégies sont évaluées sur trois corpus (scientifique, juridique et journalistique). La stratégie mixte augmente le taux de rappel (+15% sur les trois corpus

cumulés) sans toutefois modifier le taux de précision (~ 85%). Nous discutons ces performances, notamment à la lumière des résultats obtenus par ailleurs sur la préposition de.

Cet article propose une définition des arbres de dérivation pour les Grammaires d'Arbres Adjoints, étendant la notion habituelle. Elle est construite sur l'utilisation des Grammaires Catégorielles Abstraites et permet de manière symétrique le calcul de la représentation syntaxique (arbre dérivé) et le calcul de la représentation sémantique.

En recherche d'information, savoir reformuler une idée par des termes différents est une des clefs pour l'amélioration des performances des systèmes de recherche d'information (SRI) existants. L'un des moyens pour résoudre ce problème est d'utiliser des ressources sémantiques spécialisées et adaptées à la base documentaire sur laquelle les recherches sont faites. Nous proposons dans cet article de montrer que les liens sémantiques entre noms et verbes appelés liens qualia, définis dans le modèle du Lexique génératif (Pustejovsky, 1995), peuvent effectivement améliorer les résultats des SRI. Pour cela, nous extrayons automatiquement des couples nom-verbe en relation qualia de la base documentaire à l'aide du système d'acquisition ASARES (Claveau, 2003a). Ces couples sont ensuite utilisés pour étendre les requêtes d'un système de recherche. Nous montrons, à l'aide des données de la campagne d'évaluation Amaryllis, que cette extension permet effectivement d'obtenir des réponses plus pertinentes, et plus particulièrement pour les premiers documents retournés à l'utilisateur.

Les réseaux lexico-sémantiques de type Word-Net ont fait l'objet de nombreuses critiques concernant la nature des sens qu'ils distinguent ainsi que la façon dont ils caractérisent ces distinctions de sens. Cet article présente une solution possible à ces limites, solution consistant à définir les sens des mots à partir de leur usage. Plus précisément, il propose de différencier les sens d'un mot à partir d'un réseau de co-occurrences lexicales construit sur la base d'un large corpus.

Cette méthode a été testée à la fois pour le français et pour l'anglais et a fait l'objet dans ce dernier cas d'une première évaluation par comparaison avec Word-Net.

L'article présente une méthode de désambiguïsation dans laquelle le sens est déterminé en utilisant un dictionnaire. La méthode est basée sur un algorithme qui calcule une distance « sémantique » entre les mots du dictionnaire en prenant en compte la topologie complète du dictionnaire, vu comme un graphe sur ses entrées. Nous l'avons testée sur la désambiguïsation des définitions du dictionnaire elles-mêmes. L'article présente des résultats préliminaires, qui sont très encourageants pour une méthode ne nécessitant pas de corpus annoté.

L'objectif de cet article est de présenter nos travaux concernant la combinaison d'analyseurs syntaxiques pour produire un analyseur plus robuste. Nous avons créé une plate-forme nous permettant de comparer des analyseurs syntaxiques pour une langue donnée en découpant leurs résultats en informations élémentaires, en les normalisant, et en les comparant aux résultats de référence. Cette même plate-forme est utilisée pour combiner plusieurs analyseurs pour produire un analyseur de dépendance plus couvrant et plus robuste. À long terme, il sera possible de "compiler" les connaissances extraites de plusieurs analyseurs dans un analyseur de dépendance autonome.

Nous présentons dans cet article un algorithme d'apprentissage syntaxico-sémantique du langage naturel. Les données de départ sont des phrases correctes d'une langue donnée, enrichies d'informations sémantiques. Le résultat est l'ensemble des grammaires formelles satisfaisant certaines conditions et compatibles avec ces données. La stratégie employée, validée d'un point de vue théorique, est testée sur un corpus de textes français constitué pour l'occasion.

Dans cet article, nous présentons quelques résultats en catégorisation automatique de données du langage naturel sans recours à des connaissances préalables. Le système part d'une liste de

formes grammaticales françaises et en construit un graphe qui représente les chaînes rencontrées dans un corpus de textes de taille raisonnable ; les liens sont pondérés à partir de données statistiques extraites du corpus. Pour chaque chaîne de formes grammaticales significative, un vecteur reflétant sa distribution est extrait et passé à un réseau de neurones de type carte topologique auto-organisatrice. Une fois le processus d'apprentissage terminé, la carte résultante est convertie en un graphe d'étiquettes générées automatiquement, utilisé dans un tagger ou un analyseur de bas niveau. L'algorithme est aisément adaptable à toute langue dans la mesure où il ne nécessite qu'une liste de marques grammaticales et un corpus important (plus il est gros, mieux c'est). Il présente en outre un intérêt supplémentaire qui est son caractère dynamique : il est extrêmement aisé de recalculer les données à mesure que le corpus augmente.

Cet article présente une manière d'intégrer un étiqueteur morpho-syntaxique et un analyseur partiel. Cette integration permet de corriger des erreurs effectuées par l'étiqueteur seul. L'étiqueteur et l'analyseur ont été réalisés sous la forme d'automates pondérés. Des résultats sur un corpus du français ont montré une dimintion du taux d'erreur de l'ordre de 12%.

Dans le cadre du projet LIDIA, nous avons montré que dans de nombreuses situations, la TA Fondée sur le Dialogue (TAFD) pour auteur monolingue peut offrir une meilleure solution en traduction multicible que les aides aux traducteurs, ou la traduction avec révision, même si des langages contrôlés sont utilisés. Nos premières expériences ont mis en évidence le besoin de conserver les « intentions de l'auteur » au moyen « d'annotations de désambiguïsation ». Ces annotations permettent de transformer le document source en un Document Auto-Explicatif (DAE). Nous présentons ici une solution pour intégrer ces annotations dans un document XML et les rendre visibles et utilisables par un lecteur pour une meilleure compréhension du « vrai contenu » du document. Le concept de Document Auto-Explicatif pourrait changer profondément notre façon de comprendre des documents importants ou écrits dans un style complexe. Nous montrerons aussi

qu'un DAE, traduit dans une langue cible L, pourrait aussi être transformé, sans interaction humaine, en un DAE en langue L si un analyseur et un désambiguïseur sont disponibles pour cette langue L. Ainsi, un DAE pourrait être utilisé dans un contexte monolingue, mais aussi dans un contexte multi-lingue sans travail humain additionnel.

Parallèlement à l'intégration du français en TA de Parole multi-lingue (projets C-STAR, NESPOLE!), nous avons développé plusieurs plates-formes, dans le cadre des projets ERIM (Environnement Réseau pour l'Interprétariat Multimodal) et ChinFaDial (collecte de dialogues parlés spontanés français-chinois), pour traiter différents aspects de la communication orale spontanée bilingue non finalisée sur le web : interprétariat humain à distance, collecte de données, intégration d'aides automatiques (serveur de TA de Parole utilisant des composants du marché, interaction multimodale entre interlocuteurs, et prochainement aides en ligne aux intervenants, locuteurs ou interprètes). Les corpus collectés devraient être disponibles sur un site DistribDial au printemps 2004. Ces plates-formes sont en cours d'intégration, en un système générique multifonctionnel unique ERIMM d'aide à la communication multi-lingue multimodale, dont une variante s'étendra également à la formation à distance (e-training) à l'interprétariat.

Cet article présente une méthode pour extraire, à partir de corpus comparables d'un domaine de spécialité, un lexique bilingue comportant des termes simples et complexes. Cette méthode extrait d'abord les termes complexes dans chaque langue, puis les aligne à l'aide de méthodes statistiques exploitant le contexte des termes. Après avoir rappelé les difficultés que pose l'alignement des termes complexes et précisé notre approche, nous présentons le processus d'extraction de terminologies bilingues adopté et les ressources utilisées pour nos expérimentations. Enfin, nous évaluons notre approche et démontrons son intérêt en particulier pour l'alignement de termes complexes non compositionnels.

Une des conséquences du développement d'Internet et de la globalisation des échanges est le nombre considérable d'individus amenés à consulter des documents en ligne dans une langue autre que la leur. Après avoir montré que ni la traduction automatique, ni les aides terminologiques en ligne ne constituent une réponse pleinement adéquate à ce nouveau besoin, cet article présente un système d'aide à la lecture en langue étrangère basé sur un analyseur syntaxique puissant. Pour un mot sélectionné par l'usager, ce système analyse la phrase entière, de manière (i) à choisir la lecture du mot sélectionné la mieux adaptée au contexte morpho-syntaxique et (ii) à identifier une éventuelle expression idiomatique ou une collocation dont le mot serait un élément. Une démonstration de ce système, baptisé TWiC (Translation of words in context "Traduction de mots en contexte"), pourra être présentée.

Dans cet article nous présentons une application de génération de résumés multi-lingues ciblés à partir de textes d'un domaine restreint. Ces résumés sont dits ciblés car ils sont produits d'après les spécifications d'un utilisateur qui doit décider a priori du type de l'information qu'il souhaite voir apparaître dans le résumé final. Pour mener à bien cette tâche, nous effectuons dans un premier temps l'extraction de l'information spécifiée par l'utilisateur. Cette information constitue l'entrée d'un système de génération multi-lingue qui produira des résumés normalisés en trois langues (anglais, français et espagnol) à partir d'un texte en anglais.

Pour construire une ontologie, un modéliseur a besoin d'objecter des informations sémantiques sur les termes principaux de son domaine d'étude. Les outils d'exploration de corpus peuvent aider à repérer ces types d'information, et l'identification de couples d'hyperonymes a fait l'objet de plusieurs travaux. Nous proposons d'exploiter des énoncés définitoires pour extraire d'un corpus des informations concernant les trois axes de l'ossature ontologique : l'axe vertical, lié à l'hyperonymie, l'axe horizontal, lié à la co-hyponymie et l'axe transversal, lié aux relations du domaine. Après un rappel des travaux existants en repérage d'énoncés définitoires en TAL, nous

développons la méthode que nous avons mise en place, puis nous présentons son évaluation et les premiers résultats obtenus. Leur repérage atteint de 10% à 69% de précision suivant les patrons, celui des unités lexicales varie de 31% à 56%, suivant le référentiel adopté.

La publication de décisions de justice sur le Web permet de rendre la jurisprudence accessible au grand public, mais il existe des domaines du droit pour lesquels la Loi prévoit que l'identité de certaines personnes doit demeurer confidentielle. Nous développons actuellement un système d'anonymisation automatique à l'aide de l'environnement de développement GATE. Le système doit reconnaître certaines entités nommées comme les noms de personne, les lieux et les noms d'entreprise, puis déterminer automatiquement celles qui sont de nature à permettre l'identification des personnes visées par les restrictions légales à la publication.

Le Mot sur le Bout de la Langue (Tip Of the Tongue en anglais), phénomène très étudié par les psycholinguistes, nous a amené nombre d'informations concernant l'organisation du lexique mental. Un locuteur en état de TOT reconnaît instantanément le mot recherché présenté dans une liste. Il en connaît le sens, la forme, les liens avec d'autres mots... Nous présentons ici une étude de développement d'outil qui prend en compte ces spécificités, pour assister un locuteur/rédacteur à trouver le mot qu'il a sur le bout de la langue. Elle consiste à recréer le phénomène du TOT, où, dans un contexte de production un mot, connu par le système, est momentanément inaccessible. L'accès au mot se fait progressivement grâce aux informations provenant de bases de données linguistiques. Ces dernières sont essentiellement des relations de type paradigmatique et syntagmatique. Il s'avère qu'un outil, tel que SVETLAN, capable de structurer automatiquement un dictionnaire par domaine, peut être avantageusement combiné à une base de données riche en liens paradigmatiques comme EuroWord-Net, augmentant considérablement les chances de trouver le mot auquel on ne peut accéder.

Longtemps considérée comme ornementale, la structure informationnelle des documents écrits prise en charge par la morpho-disposition devient un objet d'étude à part entière dans diverses disciplines telles que la linguistique, la psycholinguistique ou l'informatique. En particulier, nous nous intéressons à l'utilité de cette dimension et, le cas échéant, son utilisabilité, dans le cadre de la transposition automatique à l'oral des textes. Dans l'objectif de fournir des solutions qui permettent de réagir efficacement à cette « inscription morphologique », nous proposons la synoptique d'un système d'oralisation. Nous avons modélisé et partiellement réalisé le module spécifique aux stratégies d'oralisation, afin de rendre « articulables » certaines parties signifiantes des textes souvent « oubliées » par les systèmes de synthèse. Les premiers résultats de cette étude ont conduit à des spécifications en cours d'intégration par un partenaire industriel. Les perspectives de ce travail peuvent intéresser la communauté TAL en reconnaissance de la parole, en génération/résumé de texte ou en multimodalité.

Cet article présente une analyse détaillée des facteurs qui déterminent les performances des approches de désambiguïsation dérivées de la méthode de Lesk (1986). Notre étude porte sur une série d'expériences concernant la méthode originelle de Lesk et des variantes que nous avons adaptées aux caractéristiques de WORDNET. Les variantes implémentées ont été évaluées sur le corpus de test de SENSEVAL2, English All Words, ainsi que sur des extraits du corpus SEMCOR. Notre évaluation se base d'un côté, sur le calcul de la précision et du rappel, selon le modèle de SENSEVAL, et d'un autre côté, sur une taxonomie des réponses qui permet de mesurer la prise de risque d'un décideur par rapport à un système de référence.

Cet article propose l'introduction d'une notion de densité syntaxique permettant de caractériser la complexité d'un énoncé et au-delà d'introduire la spécification d'un gradient de grammaticalité. Un tel gradient s'avère utile dans plusieurs cas : quantification de la difficulté d'interprétation d'une phrase, gradation de la quantité d'information syntaxique contenue dans un énoncé, explication de

la variabilité et la dépendances entre les domaines linguistiques, etc. Cette notion exploite la possibilité de caractérisation fine de l'information syntaxique en termes de contraintes : la densité est fonction des contraintes satisfaites par une réalisation pour une grammaire donnée. Les résultats de l'application de cette notion à quelques corpus sont analysés.

Les évolutions récentes des formalismes et théories linguistiques font largement appel au concept de contrainte. De plus, les caractéristiques générales des grammaires de traits ont conduit plusieurs auteurs à pointer la ressemblance existant entre ces notions et les objets ou frames. Une évolution récente de la programmation par contraintes vers les programmes de contraintes orientés objet (OOCP) possède une application possible au traitement des langages naturels. Nous proposons une traduction systématique des concepts et contraintes décrits par les grammaires de propriétés sous forme d'un OOCP. Nous détaillons l'application de cette traduction au langage "context free" archétypal anbn, en montrant que cette approche permet aussi bien l'analyse que la génération de phrases, de prendre en compte la sémantique au sein du même modèle et ne requiert pas l'utilisation d'algorithmes ad hoc pour le parsage.

Cet article propose un formalisme mathématique générique pour la combinaison de structures. Le contrôle de la saturation des structures finales est réalisé par une polarisation des objets des structures élémentaires. Ce formalisme permet de mettre en évidence et de formaliser les mécanismes procéduraux masqués de nombreux formalismes, dont les grammaires de réécriture, les grammaires de dépendance, TAG, HPSG et LFG.

Les Grammaires d'Arbres Adjoints (TAG) sont connues pour ne pas être assez puissantes pour traiter le brouillage d'arguments dans des langues à ordre des mots libre. Les variantes TAG proposées jusqu'à maintenant pour expliquer le brouillage ne sont pas entièrement satisfaisantes. Nous présentons ici une extension alternative de TAG, basée sur la notion du partage de noeuds.

En considerant des données de l'allemand et du coréen, on montre que cette extension de TAG peut en juste proportion analyser des données de brouillage d'arguments, également en combinaison avec l'extraposition et la topicalisation.

Nous nous intéressons à la construction des index de fin de livres. Nous avons développé le système IndDoc qui aide la construction de tels index. L'un des enjeux de la construction d'index est la sélection des informations : sélection des entrées les plus pertinentes et des renvois au texte les plus intéressants. Cette sélection est évidemment utile pour le lecteur qui doit trouver suffisamment d'information mais sans en être submergé. Elle est également précieuse pour l'auteur de l'index qui doit valider et corriger une ébauche d'index produite automatiquement par IndDoc. Nous montrons comment cette sélection de l'information est réalisée par IndDoc. Nous proposons une mesure qui permet de trier les entrées par ordre de pertinence décroissante et une méthode pour calculer les renvois au texte à associer à chaque entrée de l'index.

Nous présentons les résultats de l'approche statistique que nous avons développée pour le repérage de mots informatifs à partir de textes oraux. Ce travail fait partie d'un projet lancé par le département de la défense canadienne pour le développement d'un système d'extraction d'information dans le domaine de la Recherche et Sauvetage maritime (SAR). Il s'agit de trouver et annoter les mots pertinents avec des étiquettes sémantiques qui sont les concepts d'une ontologie du domaine (SAR). Notre méthode combine deux types d'information : les vecteurs de similarité générés grâce à l'ontologie du domaine et le dictionnaire-thésaurus Wordsmyth ; le contexte d'énonciation représenté par le thème. L'évaluation est effectuée en comparant la sortie du système avec les réponses de formulaires d'extraction d'information prédéfinis. Les résultats obtenus sur les textes oraux sont comparables à ceux obtenus dans le cadre de MUC7 pour des textes écrits.

Dans cet article, nous nous intéressons à la tâche de détection de thème dans le cadre de la

reconnaissance automatique de la parole. La combinaison de plusieurs méthodes de détection montre ses limites, avec des performances de 93.1 %. Ces performances nous mènent à remetttre en cause le thème de référence des paragraphes de notre corpus. Nous avons ainsi effectué une étude sur la fiabilité de ces références, en utilisant notamment les mesures Kappa et erreur de Bayes. Nous avons ainsi pu montrer que les étiquettes thématiques des paragraphes du corpus de test comportaient vraisemblablement des erreurs, les performances de détection de thème obtenues doivent donc êtres exploitées prudemment.

Cet article étudie l'adaptation au dialogue oral homme-machine des techniques de résolution des anaphores pronominales qui ont été développées par le TALN pour les documents écrits. A partir d'une étude de corpus de dialogue oral, il étudie la faisabilité de ce portage de l'écrit vers l'oral. Les résultats de cette étude montrent que certains indices utilisés à l'écrit (accord en nombre, distance entre le pronom est son antécédent) sont plus friables en dialogue oral finalisé. Les techniques développées pour l'écrit ne peuvent donc pas être réutilisées directement à l'oral.

Cet article traite de l'annotation automatique d'informations temporelles dans des textes et vise plus particulièrement les relations entre événements introduits par les verbes dans chaque clause. Si ce problème a mobilisé beaucoup de chercheurs sur le plan théorique, il reste en friche pour ce qui est de l'annotation automatique systématique (et son évaluation), même s'il existe des débuts de méthodologie pour faire réaliser la tâche par des humains. Nous proposons ici à la fois une méthode pour réaliser la tâche automatiquement et une manière de mesurer à quel degré l'objectif est atteint. Nous avons testé la faisabilité de ceci sur des dépêches d'agence avec des premiers résultats encourageants.

Cet article présente l'intégration au sein d'un analyseur syntaxique (Xerox Incremental Parser) de règles spécifiques qui permettent de lier l'analyse grammaticale à la sémantique des balises XML

spécifiques à un document donné. Ces règles sont basées sur la norme XPath qui offre une très grande finesse de description et permet de guider très précisément l'application de l'analyseur sur une famille de documents partageant une même DTD. Le résultat est alors être intégré directement comme annotation dans le document traité.

Les corpus français librement accessibles annotés à d'autres niveaux linguistiques que morpho-syntaxique sont insuffisants à la fois quantitativement et qualitativement. Partant de ce constat, la FREEBANK -- construite sur la base d'outils d'analyse automatique dont la sortie est révisée manuellement -- se veut une base de corpus du français annotés à plusieurs niveaux (structurel, morphologique, syntaxique, co-référentiel) et à différents degrés de finesse linguistique qui soit libre d'accès, codée selon des schémas normalisés, intégrant des ressources existantes et ouverte à l'enrichissement progressif.

Cet article présente l'annotation en constituants menée dans le cadre d'un protocole d'évaluation des analyseurs syntaxiques (mis au point dans le pré-projet PEAS, puis dans le projet EASY). Le choix des constituants est décrit en détail et une première évaluation effectuée à partir des résultats de deux analyseurs est donnée.

Le générateur GEPHOX que nous réalisons a pour ambition de produire des textes pour des définition ou preuves mathématiques écrites à l'aide de l'assistant de preuve PHOX. Dans cet article nous nous concentrons sur le module de détermination de contenu ContDet de GEPHOX. Après un aperçu sur l'entrée du générateur, i.e. la preuve formelle et l'ensemble des règles ayant permis de l'obtenir, nous décrivons les base de connaissances du générateur et le fonctionnement de l'algorithme de détermination de contenu.

Cet article traite de l'apprentissage symbolique de règles syntaxiques dans le modèle de Gold.

Kanazawa a montré que certaines classes de grammaires catégorielles sont apprenables dans ce modèle. L'algorithme qu'il propose nécessite une grande quantité d'information en entrée pour être efficace. En changeant la nature des informations en entrée, nous proposons un algorithme d'apprentissage de grammaires catégorielles plus réaliste dans la perspective d'applications au langage naturel.

Nous proposons d'intégrer la sémantique dans les grammaires d'interaction, formalisme qui a été conçu pour représenter la syntaxe des langues. Pour cela, nous ajoutons au formalisme un niveau supplémentaire qui s'appuie sur les mêmes principes fondamentaux que le niveau syntaxique : contrôle de la composition par un système de polarités et utilisation de la notion de description de structure pour exprimer la sous-spécification. A la différence du niveau syntaxique, les structures sont des graphes acycliques orientés et non des arbres localement ordonnés. L'interface entre les deux niveaux est assurée de façon souple par une fonction de liage qui associe à tout noeud syntaxique au plus un noeud sémantique.

Le but de cet article est de montrer pourquoi les Grammaires à Concaténation d'Intervalles (Range Concatenation Grammars, ou RCG) sont un formalisme particulièrement bien adapté à la description du langage naturel. Nous expliquons d'abord que la puissance nécessaire pour décrire le langage naturel est celle de PTIME. Ensuite, parmi les formalismes grammaticaux ayant cette puissance d'expression, nous justifions le choix des RCG. Enfin, après un aperçu de leur définition et de leurs propriétés, nous montrons comment leur utilisation comme grammaires linguistiques permet de traiter des phénomènes syntagmatiques complexes, de réaliser simultanément l'analyse syntaxique et la vérification des diverses contraintes (morpho-syntaxiques, sémantique lexicale), et de construire dynamiquement des grammaires linguistiques modulaires.

Une approche classique en recherche d'information (RI) consiste à bâtir une représentation des

documents et des requêtes basée sur les mots simples les constituant. L'utilisation de modèles bi-grammes a été étudiée, mais les contraintes sur l'ordre et l'adjacence des mots dans ces travaux ne sont pas toujours justifiées pour la recherche d'information. Nous proposons une nouvelle approche basée sur les modèles de langue qui incorporent des affinités lexicales (ALs), c'est à dire des paires non ordonnées de mots qui se trouvent proches dans un texte. Nous décrivons ce modèle et le comparons aux plus traditionnels modèles uni-grammes et bi-grammes ainsi qu'au modèle vectoriel.

Ce papier discute la Néoposie: l'inférence auto-adaptive de catégories grammaticales de mots de la langue naturelle. L'inférence grammaticale peut être divisée en deux parties : l'inférence de catégories grammaticales de mots et l'inférence de la structure. Nous examinons les éléments de base de l'apprentissage auto-adaptif du marquage des catégories grammaticales, et discutons l'adaptation des trois types principaux de marqueurs des catégories grammaticales à l'inférence auto-adaptive de catégories grammaticales de mots. Des marqueurs statistiques de n-grammes suggèrent une approche de regroupement statistique, mais le regroupement n'aide ni avec les types de mots peu fréquents, ni avec les types de mots nombreux qui peuvent se présenter dans plus d'une catégorie grammaticale. Le marqueur alternatif d'apprentissage basé sur la transformation suggère une approche basée sur la contrainte de l'unification de contextes d'occurrences de mots. Celle-ci présente un moyen de regrouper des mots peu fréquents, et permet aux occurrences différentes d'un seul type de mot d'appartenir à des catégories différentes selon les contextes grammaticaux où ils se présentent. Cependant, la simple unification de contextes d'occurrences de mots produit un nombre incroyablement grand de catégories grammaticales de mots. Nous avons essayé d'unifier plus de catégories en modérant le contexte de la correspondance pour permettre l'unification des catégories de mots aussi bien que des occurrences de mots, mais cela entraîne des unifications fausses. Nous concluons que l'avenir peut être un hybride qui comprend le regroupement de types de mots peu fréquents, l'unification de contextes d'occurrences de mots, et

le 'seeding' avec une connaissance linguistique limitée. Nous demandons un programme de nouvelles recherches pour développer une valise pour la découverte de la langue naturelle.

Après un rappel de la problématique de l'ordonnancement temporel dans un texte, nous décrivons les S-langages qui offrent une représentation unifiée des relations temporelles et une opération (la jointure) permettant de calculer les combinaisons entre celles-ci.

Le dialogue est un processus interactif pendant lequel les différents agents impliqués vont s'engager sur un certain nombre d'éléments propositionnels. La modulation implique des ajouts propositionnels - révisés et atténués - qui ne constituent pas nécessairement une base pour un accord. L'objectif de cet article est donc de proposer une description formelle du phénomène de modulation dans le cadre du modèle de J. Ginzburg.

Dans cet article, nous détaillons les résultats de la seconde évaluation du projet européen NESPOLE! auquel nous avons pris part pour le français. Dans ce projet, ainsi que dans ceux qui l'ont précédé, des techniques d'évaluation subjectives ? réalisées par des évaluateurs humains ? ont été mises en oeuvre. Nous présentons aussi les nouvelles techniques objectives ? automatiques ? proposées en traduction de l'écrit et mises en oeuvre dans le projet C-STAR III. Nous conclurons en proposant quelques idées et perspectives pour le domaine.

Cet article présente le langage de représentation des connaissances linguistiques LangTex qui permet de spécifier d'une manière unifiée les descriptions linguistiques nécessaires au repérage d'objets textuels qui organisent les textes écrits.

Nous présentons dans cet article un logiciel permettant d'assister l'usager, de manière personnalisée lors de la recherche documentaire sur le Web. L'architecture du logiciel est basée sur

l'intégration d'outils numériques de traitements des langues naturelles (TLN). Le système utilise une stratégie de traitement semi-automatique où la contribution de l'utilisateur assure la concordance entre ses attentes et les résultats obtenus.

Dans cet article, nous présentons une plate-forme de test et de recueil de dialogue oral homme-machine. Dans son architecture générale, des magiciens d'Oz simulent la compréhension des énoncés des utilisateurs et le contrôle du dialogue. Puis, nous comparons, dans un tel corpus, la prédiction statistique d'acte de dialogue avec les attentes du locuteur.

Cet article décrit une propriété sémantique propre aux dépendances syntaxiques binaires: la co-composition. On proposera ici une définition plus générale que celle donnée par Pustejovsky et que nous appelons "co-composition optionnelle". L'objet de cet article est de montrer les avantages apportées par la co-composition optionnelle dans deux tâches particulières en TAL: la désambiguïsation du sens des mots et la désambiguïsation structurale. Concernant cette deuxième tâche, nous décrirons les expériences faites sur un corpus.

Dans cet article, nous présentons le projet GÉRAF (Guide pour l'Évaluation des Résumés Automatiques Français), lequel vise l'élaboration de protocoles et la construction de corpus de résumés de référence pour l'évaluation des systèmes résumant des textes français. La finalité de ce projet est de mettre à la disposition des chercheurs les ressources ainsi créées.

Les relations transversales encodent des relations spécifiques entre les termes, par exemple localisé-dans, consomme, etc. Elles sont très souvent dépendantes des domaines, voire des corpus. Les méthodes automatiques consacrées au repérage de relations terminologiques plus classiques (hyperonymie, synonymie), peuvent générer occasionnellement les relations transversales. Mais leur repérage et typage restent sujets à une conceptualisation : ces relations ne

sont pas attendues et souvent pas connues à l'avance pour un nouveau domaine à explorer. Nous nous attachons ici à leur repérage mais surtout à leur typage. En supposant que les relations sont souvent exprimées par des verbes, nous misons sur l'étude des verbes du corpus et de leurs divers dérivés afin d'aborder plus directement la découverte des relations du domaine. Les expériences montrent que ce point d'attaque peut être intéressant, mais reste pourtant dépendant de la polysémie verbale et de la synonymie.

Dans le cadre de la recherche en sémantique lexicale, l'équipe TAL du LIRMM développe actuellement un système d'analyse des aspects thématiques des textes et de désambiguisation lexicale basé sur les vecteurs conceptuels. Pour la construction des vecteurs, les définitions provenant de sources lexicales différentes (dictionnaires à usage humain, listes de synonymes, définitions de thésaurus,) sont analysées. Aucun découpage du sens n'est présent dans la représentation : un vecteur conceptuel est associé à chaque définition et un autre pour représenter le sens global du mot. Nous souhaitons effectuer une catégorisation afin que chaque élément ne soit plus une définition mais un sens. Cette amélioration concerne bien sur directement les applications courantes (désambiguïsation, transfert lexical,) mais a aussi pour objectif majeur d'améliorer l'apprentissage de la base.

De manière générale, les linguistes informaticiens utilisent les structures de données arborescentes pour la documentation et l'analyse des données morphologiques et syntactiques. Dans cet article nous appliquons de telles structures sur des données phonologiques et nous démontrons comment de telles représentations peuvent avoir des applications utiles et pratiques en lexicographie informatique. À cet effet, nous décrivons trois modules intégrés: Le premier module définit un ensemble de caractéristiques mult-ilangages dans une structure arborescente exprimée en XML; le deuxième module parcours cet arbre et établis une généralisation sur des données contenues dans cet arborescence, optimise les données phonologiques et mets en valeur les implications des

caractéristiques. Le troisième module utilise l'information contenue dans l'arborescence comme une base de connaissance pour la génération de syllabes lexiques à caractéristiques multiples.

La gestion du but de dialogue est une tâche délicate pour le contrôleur de dialogue, car bien souvent il est en concurrence avec le gestionnaire de tâches avec lequel on le confond parfois dans certains systèmes. Dans cet article, nous présentons une stratégie dynamique de gestion de buts qui permet au contrôleur de dialogue de réduire sa dépendance au gestionnaire de tâche et lui apporte une meilleure réutilisabilité. Nous expérimentons le système dans le cadre du projet PVE (Portail Vocal d'Entreprise) dans lequel le dialogue peut se dérouler en plusieurs sessions et avec des interlocuteurs différents.

Dans le dialogue finalisé, les expressions référentielles portant sur les objets du contexte peuvent contenir des prédicats vagues ou relationnels, qu'il est difficile de traiter avec une logique propositionnelle. Inversement, les approches adaptées à ces types de prédicats sont difficilement implémentables dans un modèle générique et adaptable aux théories d'analyse linguistique. Nous proposons un modèle d'interprétation constructionnelle inspiré des grammaires de construction qui permet de modéliser le processus de résolution d'expressions référentielles extensionnelles tout en restant compatible avec la grammaire dont nous nous sommes inspirés.

Cet article présente l'influence de la zone de travail que possède une entité logicielle pour lui permettre de prédire l'état futur de son environnement, sur la constitution d'un lexique partagé par les différents membres d'une population, dans le cadre d'une variante "du jeu de désignation" (naming game).

LIKES (LInguistic and Knowledge Engineering Station) est une station d'ingénierie linguistique destinée à traiter des corpus, elle fonctionne pour l'instant sur la plupart des langues européennes

et slaves en utilisant des ressources minimales pour chaque langue. Les corpus sont constitués d'un ou plusieurs textes en ASCII ou en HTML, l'interface donne la possibilité de constituer son corpus et d'y exécuter un certain nombre de tâches allant de simples tâches de découpage en mot, de tri ou de recherche de motifs à des tâches plus complexes d'aide à la synthèse de grammaire, d'aide au repérage de relations, d'aide à la construction d'une terminologie. Nous décrivons ici les principales fonctionnalités de LIKES en rapport avec le traitement des corpus et ce qui fait sa spécificité par rapport à d'autres environnements comparables : l'utilisation minimale de ressources linguistiques.

La performance d'une résolution automatique d'anaphores infidèles pour le français pourrait atteindre une F-mesure de 30%. Ce résultat repose toutefois sur une ressource équivalente à un bon dictionnaire de la langue française, une analyse syntaxique de qualité satisfaisante et un traitement performant des entités nommées. En l'absence de telles ressources, les meilleurs résultats plafonnent autour d'une F-mesure de 15%.

Cet article présente le modèle de langage développé pour le système Sibylle, un système d'aide à la communication pour les personnes handicapées. L'utilisation d'un modèle de langage permet d'améliorer la pertinence des mots proposés en tenant compte du contexte gauche de la saisie en cours. L'originalité de notre modèle se situe dans l'intégration de la notion de chunks afin d'élargir la taille du contexte pris en compte pour l'estimation de la probabilité d'apparition des mots.

Les entitées nomées et plus généralement les multi-mots sont des ressources importantes pour plusieurs applications. Cependant, les méthodes d'extraction automatique, indépendentes de la langue, de multi-mots, ne nous donnent pas des données 100% fiables. Dans ce papier nous proposons premièrement une méthode pour selectioner entités nomées d'entre les multi-mots extraits automatiquement et, deuxièmement, une méthode de groupement des entités nomées

non-supervisionée et indépendente de la langue, en utilisant de la statistique. La deuxième phase de groupement rends l'évaluation humaine plus simple. Les traits utilisés pour le groupement sont décrits et motivés. L'analyse faite pour le groupement nous a permis d'obtenir différents groupes d'entités nomées. La méthode a été appliquée sur le bulgare et l'anglais. La précision obtenue pour certains groupes a été très haute. D'autres groupes doivent être encore rafinés. Par ailleurs, les traits discrimants appris pendant la phase de groupement nous permettent de classifier de nouvelles entités nomées.

L'apprentissage par analogie se fonde sur un principe inférentiel potentiellement pertinent pour le traitement des langues naturelles. L'utilisation de ce principe pour des tâches d'analyse linguistique présuppose toutefois une définition formelle de l'analogie entre séquences. Dans cet article, nous proposons une telle définition et montrons qu'elle donne lieu à l'implantation efficace d'un solveur d'équations analogiques sous la forme d'un transducteur fini. Munis de ces résultats, nous caractérisons empiriquement l'extension analogique de divers langages finis, correspondant à des dictionnaires de quatre langues.

Le travail présenté dans cet article a été réalisé dans le cadre d'un projet global de traduction automatique de la parole. L'approche de traduction est fondée sur un langage pivot ou Interchange Format (IF), qui représente le sens de la phrase indépendamment de la langue. Nous proposons une méthode qui intègre des informations sémantiques dans le modèle statistique de langage du système de Reconnaissance Automatique de Parole. Le principe consiste a utiliser certaines classes définies dans l'IF comme des classes sémantiques dans le modèle de langage. Ceci permet au système de reconnaissance de la parole d'analyser partiellement en IF les tours de parole. Les expérimentations realisées montrent qu'avec cette approche, le système de reconnaissance peut analyser directement en IF une partie des données de dialogues de notre application, sans faire appel au système de traduction (35% des mots ; 58% des tours de parole), tout en maintenant le

même niveau de performance du système global.

Ces travaux se basent sur l'approche computationelle et logique de Ed Stabler (?), qui donne une formalisation sous forme de grammaire du programme minimaliste de Noam Chomsky (?). La question que je veux aborder est comment, à partir d'une analyse syntaxique retrouver la forme prédicative de l'énoncé. Pour cela, il faut mettre en place une interface entre syntaxe et sémantique. C'est ce que je propose en utilisant les Grammaires Minimalistes Catégorielles (GMC) extension des GM vers le calcul de Lambeck. Ce nouveau formalisme permet une synchronisation simple avec le lambda-calcul. Parmi les questions fréquemment rencontrées dans le traitement des langues naturelles, j'interroge la performance de cette interface pour la résolution des problèmes de portée des quantificateurs. Je montre pourquoi et comment il faut utiliser un lambda-calcul plus élaboré pour obtenir les différentes lectures, en utilisant Constraint Languages for Lambda Structures -CLLS.

Nous nous intéressons à la recherche d'information en langue arabe en utilisant le modèle de l'analyse sémantique latente (LSA). Nous proposons dans cet article de montrer que le traitement linguistique et la pondération des unités lexicales influent sur la performance de la LSA pour quatre cas d'études : le premier avec un simple pré-traitement des corpus; le deuxième en utilisant un anti-dictionnaire; le troisième avec un racineur de l'arabe ; le quatrième où nous avons combiné l'anti-dictionnaire et le racineur. Globalement les résultats de nos expérimentations montrent que les traitements linguistiques ainsi que la pondération des unités lexicales utilisés améliorent la performance de LSA.

L'étiquetage des textes est un outil très important pour le traitement automatique de langage, il est utilisé dans plusieurs applications par exemple l'analyse morphologique et syntaxique des textes, l'indexation, la recherche documentaire, la voyellation pour la langue arabe, les modèles de langage

probabilistes (modèles n-classes), etc. Dans cet article nous avons élaboré un système d'étiquetage morpho-syntaxique de la langue arabe en utilisant les modèles de Markov cachés, et ceci pour construire un corpus de référence étiqueté et représentant les principales difficultés grammaticales rencontrées en langue arabe générale. Pour l'estimation des paramètres de ce modèle, nous avons utilisé un corpus d'apprentissage étiqueté manuellement en utilisant un jeu de 52 étiquettes de nature morpho-syntaxique. Ensuite on procède à une amélioration du système grâce à la procédure de réestimation des paramètres de ce modèle.

Dans le cadre du projet EpidémIA qui vise à la construction d'un système d'aide à la décision pour assister l'utilisateur dans son activité de gestion des risques sanitaires, un travail préalable sur la compositionalité des évènements (STEEL) nous a permis d'orienter notre travail dans le domaine de la localisation d'information spatio-temporelle. Nous avons construit des graphes de transducteurs pour identifier les informations temporelles sur un corpus de 100 dépêches de la langue anglaise de ProMed. Nous avons utilisé le système d'extraction d'information INTEX pour la construction de ces transducteurs. Les résultats obtenus présentent une efficacité de ces graphes pour l'identification des données temporelles.

Dans cet article, nous nous proposons de construire un lexique étiqueté selon les principes de la Linguistique Systémique Fonctionnelle (LSF) et de l'appliquer à la détection des noms de personnes ambigus dans les textes. Nous ne faisons pas d'analyse complète mais testons plutôt si certaines caractéristiques de la LSF peuvent permettre de compléter les modèles linguistiques actuellement utilisés pour la détection des entités nommées. Nous souhaitons ainsi apporter une contribution à l'application du formalisme LSF dans l'analyse automatique de textes après son application déjà éprouvée à la génération de textes.

Dans ce papier, nous présentons les résultats d'une étude expérimentale de la durée des

consonnes géminées de l'arabe. Nous visons à déterminer la durée, pour une séquence VCCV, de la consonne géminée CC ainsi que de la voyelle qui la précède. Nous comparons ces valeurs à celles mesurées pour une séquence VCV. Les résultats ont prouvé que la durée de la consonne simple était sensiblement différente de celle géminée, ainsi que la durée de la voyelle précédant la consonne. A la base, ce travail est entrepris dans un but d'étudier l'utilisation des durées de phonèmes comme une source d'information pour optimiser un système de reconnaissance, donc introduire des modèles explicites de durée des phonèmes, et mettre en application ces modèles comme partie du modèle acoustique du système de reconnaissance.

Alors que de nombreux travaux portent actuellement sur la linguistique de corpus, l'utilisation de textes authentiques en classe de langue, ou de corpus dans l'enseignement des langues (via concordanciers), quasiment aucun travail n'a été réalisé en vue de la réalisation de bases de textes à l'usage des enseignants de langue, indexées en fonction de critères relevant de la problématique de la didactique des langues. Dans le cadre de cet article, nous proposons de préciser cette notion d'indexation pédagogique, puis de présenter les principaux standards de description de ressources pédagogiques existants, avant de montrer l'inadéquation de ces standards à la description de textes dans l'optique de leur utilisation dans l'enseignement des langues. Enfin nous en aborderons les conséquences relativement à la réalisation de la base.

Nous exposons ici une méthode permettant d'étudier la nature d'un signal de parole dans le temps. Plus précisément, nous nous intéressons à la caractéristique de nasalisation du signal. Ainsi nous cherchons à savoir si à un instant t le signal est nasalisé ou oralisé. Nous procédons par classification à l'aide d'un réseau de neurones type perceptron multi-couches, après une phase d'apprentissage supervisée. La classification, après segmentation du signal en fenêtres, nous permet d'associer à chaque fenêtre de signal une étiquette renseignant sur la nature du signal.

Cet article présente une méthode d'indexation automatique de documents basée sur une approche linguistique et statistique. Cette dernière est une combinaison séquentielle de l'analyse linguistique du document à indexer par l'extraction des termes significatifs du document et de l'analyse statistique par la décomposition en valeurs singulières des mots composant le document. La pondération des termes tire avantage de leur contexte local, par rapport au document, global, par rapport à la base de données, et de leur position par rapport aux autres termes, les co-occurrences. Le système d'indexation présenté fait des propositions d'affectations du document à un référentiel métier dont les thèmes sont prédéfinis. Nous présentons les résultats de l'expérimentation de ce système menée sur un corpus des pôles métiers de la société Suez-Environnement.

Notre article s'intègre dans le cadre du projet intitulé "Oréodule" : un système embarqué temps réel de reconnaissance, de traduction et de synthèse de la parole. L'objet de notre intérêt dans cet article est la présentation de notre système de synthèse hybride de la parole arabe. Nous présenterons, dans ce papier, les différents modules et les différents choix techniques de notre système de synthèse hybride par concaténation de polyphèmes. Nous détaillerons également les règles de transcription et leurs effets sur le traitement linguistique, les règles de syllabation et leurs impacts sur le coût (temps et difficulté) de réalisation du module acoustique et nous poursuivrons par l'exposé de nos choix au niveau du module de concaténation. Nous décrirons le module de lissage, un traitement acoustique, post concaténation, nécessaire à l'amélioration de la qualité de la voix synthétisée. Enfin, nous présenterons les résultats de l'étude statistique de compréhension, réalisée sur un corpus.

Le Web a causé beaucoup de changements dans plusieurs domaines. Il a aussi influencé l'inventaire des genres textuels traditionnels. De nouveaux genres ont été créés, par exemple les blogues et foires aux questions. Il est probable que d'autres genres soient en train de se former, parce que le Web est un médium qui change constamment. Dans cet article, nous présentons une

expérience qui vise à faire apparaître de façon inductive les plans textuels émergents, qui peuvent devenir un nouveau genre ou une nouvelle typologie textuelle dans peu de temps. Il s'agit de regrouper (analyse de groupement) les pages web en utilisant des traits linguistiques et de présentation. Les résultats sont encourageants et invitent à poursuivre la recherche dans ce domaine.

Jusqu'a présent il n'y a pas de système automatique complet pour l'étiquetage du texte arabe. Les méthodes qu'elles soient basées sur des règles explicites ou sur des calculs statistiques, ont été développées pour pallier au problème de l'ambiguïté lexicale. Celles-ci introduisent des informations sur le contexte immédiat des mots, mais font l'impasse sur les exceptions qui échappent aux traitements. L'apparition des méthodes Memory-Based Learning(MBL) a permis l'exploitation automatique de la similarité de l'information contenue dans de grandes masses de textes et , en cas d'anomalie, permet de déduire la catégorie la plus probable dans un contexte donné, sans que le linguiste ait à formuler des règles explicites. Ce papier qui présente une approche hybride combine les méthodes à base de règles et MBL afin d'optimiser la performance de l'étiqueteur. Les résultats ainsi obtenus, présentés en section 6, sont satisfaisants et l' objectif recherché est atteint.

A partir du concept de cohésion comme mesure de l'unité du texte et du modèle oulipien de la littérature par contraintes, notre étude propose une méthode d'analyse potentielle sur ordinateur dans le cas des Cent mille milliards des poèmes. En s'appuyant sur un ensemble de contraintes initiales, notre programme serait capable d'analyser tous les textes potentiels produits par la machine en utilisant ces contraintes.

Dans le cadre du développement des environnements d'analyse linguistique, d'étiquetage de corpus et d'analyse statistique afin de traiter des corpus de grande taille, nous proposons de mettre au point des procédures nouvelles d'étiquetage morpho-syntaxique et sémantique. Nous présentons un

ensemble de ressources linguistiques - dictionnaires et grammaires - dans le but d'étiqueter entièrement le roman proustien : « Du côté de chez Swann ». Notre recherche avance deux atouts majeurs : la précision des étiquettes attribuées aux formes linguistiques du texte ; et le repérage et étiquetage exhaustifs des mots composés.

Notre article s'intègre dans le cadre du projet intitulé Oréodule: un système de reconnaissance, de traduction et de synthèse de la parole spontanée. L'objectif de cet article est de présenter un modèle d'étiquetage probabiliste, selon une approche componentielle et sélective. Cette approche ne considère que les éléments de l'énoncé porteurs de sens. La signification de chaque mot est représentée par un ensemble de traits sémantiques Ts. Ce modèle participe au choix des Ts candidats lors du décodage sémantique d'un énoncé.

Au cours des dix dernières années, l'analyse de la sémantique latente (LSA) a été utilisée dans de nombreuses approches TAL avec parfois de remarquables succès. Cependant, ses capacités à exprimer des ressemblances sémantiques n'ont pas été réellement recherchées de façon systématique. C'est l'objectif de ce travail, où la LSA est appliquée à un corpus de textes de langue courante (journal allemand). Les relations lexicales entre un mot et ses termes les plus proches sont analysés pour un test de vocabulaire. Ces résultats sont alors comparés avec les résultats obtenus lors d'une analyse des collocations.

Les systèmes de Questions Réponse ont besoin de connaissances sémantiques pour trouver dans les documents des termes susceptibles d'être des reformulations des termes de la question. Cependant, l'utilisation de ressources sémantiques peut apporter un bruit important et altérer la précision du système. ne fournit qu'une partie des reformulations possibles. Cet article présente un cadre d'évaluation pour les ressources sémantiques dans les systèmes de question-réponse. Il décrit la fabrication semi-automatique d'un corpus de questions et de réponses destiné à étudier les

reformulations présentes entre termes de la question et termes de la réponse. Il étudie la fréquence et la fiabilité des reformulations extraites de l'ontologie WordNet.

Cet article présente les principes de fonctionnement et les intérêts d'une plate-forme logicielle centrée sur un utilisateur ou un groupe d'utilisateurs et dédiée à la visualisation de propriétés thématiques d'ensembles de documents électroniques. Cette plate-forme, appelée ProxiDocs, permet de dresser des représentations graphiques (des cartes) d'un ensemble de textes à partir de thèmes choisis et définis par un utilisateur ou un groupe d'utilisateurs. Ces cartes sont interactives et permettent de visualiser les proximités et les différences thématiques entre textes composant le corpus étudié. Selon le type d'analyse souhaitée par l'utilisateur, ces cartes peuvent également s'animer afin de représenter les changements thématiques d'un ensemble de textes au fil du temps.

Nous décrivons une méthode de segmentation morphologique automatique. L'algorithme utilise uniquement une liste des mots d'un corpus et tire parti des probabilités conditionnelles observées entre les sous-chaînes extraites de ce lexique. La méthode est également fondée sur l'utilisation de graphes d'alignement de segments de mots. Le résultat est un découpage de chaque mot sous la forme (préfixe*) + base + (suffixe*). Nous évaluons la pertinence des familles morphologiques découvertes par l'algorithme sur un corpus de textes médicaux français contenant des mots à la structure morphologique complexe.

Cet article propose d'exploiter les similitudes constructionnelles de deux langues morphologiquement proches (le français et l'italien), pour créer des règles de construction des mots capables de déconstruire un néologisme construit de la langue source et générer de manière similaire un néologisme construit dans la langue cible. Nous commençons par présenter diverses motivations à cette méthode, puis détaillons une expérience pour laquelle plusieurs règles de transfert ont été créées et appliquées à un ensemble de néologismes construits.

Cet article décrit l'analyse sémantique des spécificités dans le domaine technique des machines-outils pour l'usinage des métaux. Le but de cette étude est de vérifier si et dans quelle mesure les spécificités dans ce domaine sont monosémiques ou polysémiques. Les spécificités (situées dans un continuum de spécificité) seront identifiées avec la KeyWords Method en comparant le corpus d'analyse à un corpus de référence. Elles feront ensuite l'objet d'une analyse sémantique automatisée à partir du recouvrement des co-occurrences des co-occurrences, afin d'établir le continuum de monosémie. Les travaux de recherche étant en cours, nous présenterons des résultats préliminaires de cette double analyse.

Le présent article décrit le Système AlALeR (Système d'Alignement Autonome, Léger et Robuste). Capable d'aligner au niveau phrastique un texte en français et un texte en japonais, le Système AlALeR ne recourt cependant à aucun moyen extérieur tel qu'un analyseur morphologique ou des dictionnaires, au contraire des méthodes existantes. Il est caractérisé par son analyse morphologique partielle mettant à profit des particularités du système d'écriture japonais et par la transcription des mots emprunts, à l'aide d'un transducteur.

Cet article présente une méthode d'acquisition semi-automatique de relations lexicales bilingues (français-anglais) faisant appel à un processus de validation sur le Web. Notre approche consiste d'abord à extraire automatiquement des relations lexicales françaises. Nous générons ensuite leurs traductions potentielles grâce à un dictionnaire électronique. Ces traductions sont enfin automatiquement filtrées à partir de requêtes lancées sur le moteur de recherche Google. Notre évaluation sur 10 mots français très polysémiques montre que le Web permet de constituer ou compléter des bases de données lexicales multilingues, encore trop rares, mais dont l'utilité est pourtant primordiale pour de nombreuses applications, dont la traduction automatique.

Dans cet article nous décrivons le développement des ressources linguistiques du finnois pour un système de traduction automatique de la parole dans le domaine médical: MedSLT. Le travail inclut la construction des corpus médicaux en finnois, le développement de la grammaire finlandaise pour la génération, le développement du lexique finlandais et la définition des règles de mapping interlingue-finnois pour la traduction multilingue. Nous avons découvert que le finnois peut être introduit dans l'architecture existante de MedSLT sans trop de difficultés. En effet, malgré les différences entre l'anglais et le finnois, la grammaire finlandaise a pu être créée en adaptant manuellement la grammaire anglaise originale. Les premiers résultats de l'évaluation de la traduction anglais-finnois sont encourageants.

Nous présentons dans cet article un corpus de français tchaté, destiné à l'étude de la langue du tchat. Ce corpus, collecté et encodé automatiquement, est remarquable avant tout par son étendue, puisqu'il couvre un total de 4 millions de messages sur 105 canaux, hétérogènes sur les plans thématique et pragmatique. Son codage simple ne sera toutefois pas satisfaisant pour tous les usages. Il est disponible sur un site Internet, et consultable grâce à une interface web.

Cet article présente une étude dont l'objectif était d'améliorer la phonétisation d'un système de synthèse vocale de SMS en ce qui concerne trois types de problèmes : l'écriture rébus (chiffres et lettres utilisés pour leur valeur phonique), les abréviations sous forme de squelettes consonantiques et les agglutinations (déterminants ou pronoms collés graphiquement au mot qui suit). Notre approche se base sur l'analyse d'un corpus de SMS, à partir duquel nous avons extrait des listes de formes permettant de compléter les lexiques du système, et mis au point de nouvelles règles pour les grammaires internes. Les modifications effectuées apportent une amélioration substantielle du système, bien qu'il reste, évidemment, de nombreuses autres classes de problèmes à traiter.

Dans le contexte de l'étiquetage morpho-syntaxique des corpus de spécialité, nous proposons une

approche inductive pour réduire les erreurs les plus difficiles et qui persistent après étiquetage par le système de Brill. Nous avons appliqué notre système sur deux types de confusions. La première confusion concerne un mot qui peut avoir les étiquettes ?verbe au participe passé', ?verbe au passé' ou ?adjectif'. La deuxième confusion se produit entre un nom commun au pluriel et un verbe au présent, à la 3ème personne du singulier. A l'aide d'interface conviviale, l'expert corrige l'étiquette du mot ambigu. A partir des exemples annotés, nous induisons des règles de correction. Afin de réduire le coût d'annotation, nous avons utilisé l'apprentissage actif. La validation expérimentale a montré une amélioration de la précision de l'étiquetage. De plus, à partir de l'annotation du tiers du nombre d'exemples, le niveau de précision réalisé est équivalent à celui obtenu en annotant tous les exemples.

Cet article a pour objet le métalangage définitionnel de la base de données lexicale BDéf, plus précisément l'utilisation de ce métalangage dans la modélisation des structures polysémiques du français. La Bdéf encode sous forme de définitions lexicographiques les sens lexicaux d'un sous-ensemble représentatif du lexique du français parmi lequel on compte environ 500 unités polysémiques appartenant aux principales parties du discours. L'article comprend deux sections. La première présente le métalangage de la BDéf et le situe par rapport aux différents types de définitions lexicales, qu'elles soient ou non formelles, qu'elles visent ou non l'informatisation. La seconde section présente une application de la BDéf qui vise à terme à rendre compte de la polysémie régulière du français. On y présente, à partir d'un cas spécifique, la notion de patron de polysémie.

Nous présentons les résultats de notre approche d'apprentissage de relations prédicat-argument dans le but de générer des patrons d'extraction pour des textes conversationnels. Notre approche s'effectue en trois étapes incluant la segmentation linguistique des textes pour définir des unités linguistiques à l'instar de la phrase pour les textes bien formés tels que les dépêches

journalistiques. Cette étape prend en considération la dimension discursive importante dans ces types de textes. La deuxième étape effectue la résolution des anaphores pronominales en position de sujet. Cela tient compte d'une particularité importante des textes conversationnels : la pronominalisation du thème. Nous montrons que la résolution d'un sous ensemble d'anaphores pronominales améliore l'apprentissage des patrons d'extraction. La troisième utilise des modèles de Markov pour modéliser les séquences de classes de mots et leurs rôles pour un ensemble de relations données. Notre approche expérimentée sur des transcriptions de conversations téléphoniques dans le domaine de la recherche et sauvetage identifie les patrons d'extraction avec un F-score moyen de 73,75 %.

Dans cet article, nous proposons un nouvel analyseur syntaxique, qui repose sur une variante du modèle Lexical-Functional Grammars (Grammaires Lexicales Fonctionnelles) ou LFG. Cet analyseur LFG accepte en entrée un treillis de mots et calcule ses structures fonctionnelles sur une forêt partagée. Nous présentons également les différentes techniques de rattrapage d'erreurs que nous avons mises en oeuvre. Puis nous évaluons cet analyseur sur une grammaire à large couverture du français dans le cadre d'une utilisation à grande échelle sur corpus variés. Nous montrons que cet analyseur est à la fois efficace et robuste.

Nous présentons un étiqueteur morpho-syntaxique du français. Celui-ci utilise l'apprentissage supervisé à travers un modèle de Markov caché. Le modèle de langage est appris à partir d'un corpus étiqueté. Nous décrivons son fonctionnement et la méthode d'apprentissage. L'étiqueteur atteint un score de précision de 89 % avec un jeu d'étiquettes très riche. Nous présentons ensuite des résultats détaillés pour chaque classe grammaticale et étudions en particulier la reconnaissance des homographes.

Cet article s'intéresse a la désambiguïsation sémantique d'unités lexicales alignées a travers un

corpus multi-lingue. Nous appliquons une méthode automatique non supervisée basée sur la comparaison de réseaux sémantiques, et nous dégageons un critère permettant de déterminer a priori si 2 unités alignées ont une chance de se désambiguïser mutuellement. Enfin, nous développons une méthode fondée sur un apprentissage a partir de contextes bilingues. En appliquant ce critère afin de déterminer pour quelles unités l'information traductionnelle doit être prise en compte, nous obtenons une amélioration des résultats.

Dans cet article nous présentons un langage de navigation textuelle et son implantation dans la plate-forme Navitexte. Nous décrivons une application de ces principes de navigation dans un cadre d'apprentissage de la bonne formation des textes, destinée à des dans un cadre d'apprentissage de la bonne formation des textes, destinée à des étudiants apprenant le français langue étrangère.

Cet article apporte une méthode de développement grammatical pour la réalisation de grammaires d'arbres adjoints (TAG) de taille importante augmentées d'une dimension sémantique. La méthode que nous présentons s'exprime dans un langage informatique de représentation grammatical qui est déclaratif et monot-one. Pour arriver au résultat, nous montrons comment tirer parti de la théorie de la projection dans le langage de représentation que nous utilisons. Par conséquent cet article justifie l'utilisation d'un langage monot-one pour la représentation lexico-grammaticale.

Cet article introduit une méthodologie d'intégration de la personnalité dans un système de dialogue automatique, en vue de l'incarnation de personnages virtuels. Notion complexe non encore épuisée dans la littérature, la personnalité d'un individu peut s'illustrer de multiples manières possibles. Notre objectif consiste à présenter une méthode générique de prise en compte de la personnalité dans un système de dialogue par modélisation et exploitation des connaissances relatives à la personnalité de l'individu à incarner. Cet article présente les avantages et inconvénients de cette méthode en l'illustrant au travers de la stylistique des énoncés générés par le système.

L'objectif du projet RITEL est de réaliser un système de dialogue homme-machine permettant à un utilisateur de poser oralement des questions, et de dialoguer avec un système de recherche d'information généraliste (par exemple, chercher sur l'Internet "Qui est le Président du Sénat ?") et d'en étudier les potentialités. Actuellement, la plateforme RITEL permet de collecter des corpus de dialogue homme-machine. Les utilisateurs peuvent parfois obtenir une réponse, de type factuel (Q : qui est le président de la France ; R : Jacques Chirac.). Cet article présente brièvement la plateforme développée, le corpus collecté ainsi que les questions que soulèvent un tel système et quelques unes des premières solutions envisagées.

L'objectif de cet article est la présentation d'un système de génération automatique de dictionnaires électroniques de la langue arabe classique, développé au sein du laboratoire RIADI (unité de Monastir). Ce système entre dans le cadre du projet "oreillodule": un système embarqué de synthèse, traduction et reconnaissance de la parole arabe. Dans cet article, nous présenterons, les différentes étapes de réalisation, et notamment la génération automatique de ces dictionnaires se basant sur une théorie originale : les Conditions de Structures Morphématiques (CSM), et les matrices lexicales.

Nous proposons dans cet article une approche de segmentation de textes arabes non voyellés basée sur une analyse contextuelle des signes de ponctuations et de certaines particules, tels que les conjonctions de coordination. Nous présentons ensuite notre système STAr, un segmenteur de textes arabes basé sur l'approche proposée. STAr accepte en entrée un texte arabe en format txt et génère en sortie un texte segmenté en paragraphes et en phrases.

Il a été montré que les Grammaires d'Arbres Adjoints Ensemblistes (Multicomponent Tree Adjoining Grammars, MCTAG) sont très utiles pour des applications TAL. Pourtant, la définition des MCTAG

est problématique parce qu'elle fait référence au procès de dérivation même : une contrainte de simultanéité est imposée concernant la façon dont on ajoute les membres d'un même ensemble d'arbres. En regardant uniquement le résultat d'une dérivation, c'est-à-dire l'arbre dérivé et l'arbre de dérivation, cette simultanéité n'est plus visible. Par conséquent pour vérifier la contrainte de simultanéité, il faut toujours considérer l'ordre concret des pas de la dérivation. Afin d'éviter cela, nous proposons une caractérisation alternative de MCTAG qui permet une abstraction de l'ordre de dérivation : Les arbres générés par la grammaire sont caractérisés par les propriétés de leurs arbres de dérivation.

La traduction automatique (TA) attire depuis plusieurs années l'intérêt d'un nombre grandissant de chercheurs. De nombreuses approches sont proposées et plusieurs campagnes d'évaluation rythment les avancées faites. La tâche de traduction à laquelle les participants de ces campagnes se prêtent consiste presque invariablement à traduire des articles journalistiques d'une langue étrangère vers l'anglais; tâche qui peut sembler artificielle. Dans cette étude, nous nous intéressons à savoir ce que différentes approches basées sur les corpus peuvent faire sur une tâche réelle. Nous avons reconstruit à cet effet l'un des plus grands succès de la TA: le système MÉTÉO. Nous montrons qu'une combinaison de mémoire de traduction et d'approches statistiques permet d'obtenir des résultats comparables à celles du système MÉTÉO, tout en offrant un cycle de développement plus court et de plus grandes possibilités d'ajustements.

Cet article traite du problème de la compréhensibilité des textes et en particulier du besoin de simplifier la complexité syntaxique des phrases pour des lecteurs souffrant de troubles de la compréhension. Nous présentons une approche à base de règles de simplification développées manuellement et son intégration dans un traitement de texte. Cette intégration permet la validation interactive de simplifications candidates produites par le système, et lie la tâche de création de texte simplifié à celle de rédaction.

Depuis quelques années, médecins et documentalistes doivent faire face à une demande croissante dans le domaine du codage médico-économique et de l'indexation des diverses sources d'information disponibles dans le domaine de la santé. Il est donc nécessaire de développer des outils d'indexation automatique qui réduisent les délais d'indexation et facilitent l'accès aux ressources médicales. Nous proposons deux méthodes d'indexation automatique de ressources de santé à l'aide de paires de descripteurs MeSH. La combinaison de ces deux méthodes permet d'optimiser les résulats en exploitant la complémentarité des approches. Les performances obtenues sont équivalentes à celles des outils de la littérature pour une indexation à l'aide de descripteurs seuls.

Dans cet article, un environnement modulaire pour la simulation automatique de dialogues homme-machine est proposé. Cet environnement comprend notamment un modèle d'utilisateur consistant dirigé par le but et un module de simulation de compréhension de parole. Un réseau bayésien est à la base de ces deux modèles et selon les paramètres utilisés, il peut générer un comportement d'utilisateur cohérent ou servir de classificateur de concepts. L'environnement a été utilisé dans le contexte de l'optimisation de stratégies de dialogue sur une tâche simple de remplissage de formulaire et les résultats montrent qu'il est alors possible d'identifier certains dialogues problématiques du point de vue de la compréhension.

Cet article expose un cas concret d'utilisation d'une grammaire de contraintes. Le produit qui les applique a été commercialisé en 2003 pour corriger automatiquement et en temps réel les fautes d'accord présentes dans les sous-titres des retransmissions en direct des débats du Sénat du Canada. Avant la mise en place du système, le taux moyen de fautes était de l'ordre de 7 pour 100 mots. Depuis la mise en service, le taux d'erreurs a chuté à 1,7 %. Nous expliquons dans ce qui suit les principaux atouts des grammaires de contraintes dans le cas particulier des traitements temps

réel, et plus généralement pour toutes les applications qui nécessitent une analyse au fur et à mesure du discours (c.-à-d. sans attendre la fin des phrases).

Nous présentons dans cet article un nouveau formalisme linguistique qui repose sur les Grammaires à Concaténation d'Intervalles (RCG), appelé Méta-RCG. Nous exposons tout d'abord pourquoi la non-linéarité permet une représentation adéquate des phénomènes linguistiques, et en particulier de l'interaction entre les différents niveaux de description. Puis nous présentons les Méta-RCG et les concepts linguistiques supplémentaires qu'elles mettent en oeuvre, tout en restant convertibles en RCG classiques. Nous montrons que les analyses classiques (constituants, dépendances, topologie, sémantique prédicat-arguments) peuvent être obtenues par projection partielle d'une analyse Méta-RCG complète. Enfin, nous décrivons la grammaire du français que nous développons dans ce nouveau formalisme et l'analyseur efficace qui en découle. Nous illustrons alors la notion de projection partielle sur un exemple.

Dans cet article, nous avons examiné la relation entre pause et ponctuation (virgule, point et virgule, deux-points). Toutes ces pauses sont internes aux phrases. À l'aide de l'analyse de plusieurs milliers de pauses dans un corpus de presque 17 heures d'enregistrement réalisé par une locutrice professionnelle native du portugais brésilien, nous avons vérifié une proportion importante des pauses hors ponctuations (61,3%). Les données renforcent aussi la présence des structures topique/commentaire dans la lecture à haute voix. Les résultats des durées de pause correspondantes aux ponctuations sont consistants avec les données présentées dans les grammaires.

Cet article propose une méthode innovante et efficace pour segmenter un texte en parties thématiquement cohérentes, en utilisant des chaînes lexicales pondérées. Les chaînes lexicales sont construites en fonction de hiatus variables, ou bien sans hiatus, ou encore pondérées en

fonction de la densité des occurrences du terme dans la chaîne. D'autre part, nous avons constaté que la prise en compte du repérage d'entités nommées dans la chaîne de traitement, du moins sans résolution des anaphores, n'améliore pas significativement les performances. Enfin, la qualité de la segmentation proposée est stable sur différentes thématiques, ce qui montre une indépendance par rapport au type de document.

Nous présentons une plateforme de développement de lexique offrant une base lexicale accompagnée d'un certain nombre d'outils de maintenance et d'utilisation. Cette base, qui comporte aujourd'hui 440.000 formes du Français contemporain, est destinée à être diffusée et remise à jour régulièrement. Nous exposons d'abord les outils et les techniques employées pour sa constitution et son enrichissement, notamment la technique de calcul des fréquences lexicales par catégorie morpho-syntaxique. Nous décrivons ensuite différentes approches pour constituer un sous-lexique de taille réduite, dont la particularité est de couvrir plus de 90% de l'usage. Un tel lexique noyau offre en outre la possibilité d'être réellement complété manuellement avec des informations sémantiques, de valence, pragmatiques etc.

À travers la présentation de la plate-forme LinguaStream, nous présentons certains principes méthodologiques et différents modèles d'analyse pouvant permettre l'articulation de traitements sur corpus. Nous envisageons en particulier les besoins nés de perspectives émergentes en TAL telles que l'analyse du discours.

Cet article présente un environnement de développement pour les méta-grammaires (MG), utilisé pour concevoir rapidement une grammaire d'arbres adjoints (TAG) du français à large couverture et néanmoins très compacte, grâce à des factorisations d'arbres. Exploitant les fonctionnalités fournies par le système DYALOG, cette grammaire a permis de construire un analyseur syntaxique hybride TAG/TIG utilisé dans le cadre de la campagne d'évaluation syntaxique EASY.

Dans cet article, nous présentons un outil permettant de produire automatiquement des ressources linguistiques, en l'occurence des grammaires. Cet outil se caractérise par son extensibilité, tant du point de vue des formalismes grammaticaux supportés (grammaires d'arbres adjoints et grammaires d'interaction à l'heure actuelle), que de son architecture modulaire, qui facilite l'intégration de nouveaux modules ayant pour but de vérifier la validité des structures produites. En outre, cet outil offre un support adapté au développement de grammaires à portée sémantique.

L'objectif de cet article est de présenter l'état actuel du modèle de la Grammaire d'Unification Sens-Texte, notamment depuis que les bases formelles du modèle ont été éclaircies grâce au développement des Grammaires d'Unification Polarisées. L'accent est mis sur l'architecture du modèle et le rôle de la polarisation dans l'articulation des différents modules ? l'interface sémantique-syntaxe, l'interface syntaxe-morphotopologie et les grammaires décrivant les différents niveaux de représentation. Nous étudions comment les procédures d'analyse et de génération sont contrôlables par différentes stratégies de neutralisation des différentes polarités.

Plusieurs travaux antérieurs ont fait état de l'amélioration possible des performances des systèmes de recherche documentaire grace à l'utilisation d'indexation sémantique utilisant une ontologie (p.ex. Word-Net). La présente contribution décrit une nouvelle méthode visant à réduire le nombre de termes d'indexation utilisés dans une indexation sémantique, en cherchant la coupe de redondance minimale dans la hiérarchie fournie par l'ontologie. Les résultats, obtenus sur diverses collections de documents en utilisant le dictionnaire EDR, sont présentés.

Les systèmes de questions-réponses, essentiellement focalisés sur des questions factuelles en domaine ouvert, testent également d'autres tâches, comme le travail en domaine contraint ou la recherche de définitions. Nous nous intéressons ici à la recherche de réponses à des questions «

définitoires » portant sur le domaine médical. La recherche de réponses de type définitoire se fait généralement en utilisant deux types de méthodes : celles s'appuyant essentiellement sur le contenu du corpus cible, et celles faisant appel à des connaissances externes. Nous avons choisi de nous limiter au premier de ces deux types de méthodes. Nous présentons une expérience dans laquelle nous réutilisons des patrons de repérage d'énoncés définitoires, conçus pour une autre tâche, pour localiser les réponses potentielles aux questions posées. Nous avons intégré ces patrons dans une chaîne de traitement que nous évaluons sur les questions définitoires et le corpus médical du projet EQueR sur l'évaluation de systèmes de questions-réponses. Cette évaluation montre que, si le rappel reste à améliorer, la « précision » des réponses obtenue (mesurée par la moyenne des inverses de rangs) est honorable. Nous discutons ces résultats et proposons des pistes d'amélioration.

QRISTAL (Questions-Réponses Intégrant un Système de Traitement Automatique des Langues) est un système de questions-réponses utilisant massivement le TAL, tant pour l'indexation des documents que pour l'extraction des réponses. Ce système s'est récemment classé premier lors de l'évaluation EQueR (Evalda, Technolanguez). Après une description fonctionnelle du système, ses performances sont détaillées. Ces résultats et des tests complémentaires permettent de mieux situer l'apport des différents modules de TAL. Les réactions des premiers utilisateurs incitent enfin à une réflexion sur l'ergonomie et les contraintes des systèmes de questions-réponses, face aux outils de recherche sur le Web.

Cet article s'intéresse à la manière dont la morphosémantique peut contribuer à l'appariement multi-lingue de variantes terminologiques entre termes. L'approche décrite permet de relier automatiquement entre eux les noms et adjectifs composés savants d'un corpus spécialisé en médecine (synonymie, hyponymie, approximation). L'acquisition de relations lexicales est une question particulièrement cruciale lors de l'élaboration de bases de données et de systèmes de

recherche d'information multi-lingues. La méthode est applicable à au moins cinq langues européennes dont elle exploite les caractéristiques morphologiques similaires des mots composés dans les langues de spécialité. Elle consiste en l'intéraction de trois dispositifs : (1) un analyseur morphosémantique monolingue, (2) une table multi-lingue qui définit des relations de base entre les racines gréco-latines des lexèmes savants, (3) quatre règles indépendantes de la langue qui infèrent, à partir de ces relations de base, les relations lexicales entre les lexèmes contenant ces racines. L'approche décrite est implémentée en français, où l'on dispose d'un analyseur morphologique capable de calculer la définition de mots construits inconnus à partir du sens de ses composants. Le corpus de travail est un lexique spécialisé médical d'environ 29000 lexèmes, que le calcul des relations de synonymie, hyponymie et approximation a permis de regrouper en plus de 3000 familles lexicales.

Dans le cadre de la recherche sur la représentation du sens en Traitement Automatique des Langues Naturelles, nous nous concentrons sur la construction d'un système capable d'acquérir le sens des mots, et les relations entre ces sens, à partir de dictionnaires à usage humain, du Web ou d'autres ressources lexicales. Pour l'antonymie, il n'existe pas de listes séparant les antonymies complémentaire, scalaire et duale. Nous présentons dans cet article une approche semi-supervisée permettant de construire ces listes. Notre méthode est basée sur les oppositions de nature morphologique qui peuvent exister entre les items lexicaux. À partir d'un premier ensemble de couples antonymes, elle permet non seulement de construire ces listes mais aussi de trouver des oppositions morphologiques. Nous étudions les résultats obtenus par cette méthode. En particulier, nous présentons les oppositions de préfixes ainsi découvertes et leur validité sur le corpus puis nous discutons de la répartition des types d'antonymie en fonction des couples opposés de préfixes.

Les ressources linguistiques les plus facilement disponibles en TAL ressortissent généralement au registre général d'une langue. Lorsqu'elles doivent être utilisées sur des textes de spécialité il peut

être utile de les adapter à ces textes. Cet article est consacré à l'adaptation de ressources synonymiques générales à la langue médicale. L'adaptation est obtenue suite à une série de filtrages sur un corpus du domaine. Les synonymes originaux et les synonymes filtrés sont ensuite utilisés comme une des ressources pour la normalisation de variantes de termes dans une tâche de structuration de terminologie. Leurs apports respectifs sont évalués par rapport à la structure terminologique de référence. Cette évaluation montre que les résultats sont globalement encourageants après les filtrages, pour une tâche comme la structuration de terminologies : une amélioration de la précision contre une légère diminution du rappel.

L'analyse syntaxique reste un problème complexe au point que nombre d'applications n'ont recours qu'à des analyseurs superficiels. Nous faisons dans cet article le point sur les notions d'analyse superficielles et profondes en proposant une première caractérisation de la notion de complexité opérationnelle pour l'analyse syntaxique automatique permettant de distinguer objets et relations plus ou moins difficiles à identifier. Sur cette base, nous proposons un bilan des différentes techniques permettant de caractériser et combiner analyse superficielle et profonde.

Cet article expose l'ensemble des outils que nous avons mis en oeuvre pour la campagne EASy d'évaluation d'analyse syntaxique. Nous commençons par un aperçu du lexique morphologique et syntaxique utilisé. Puis nous décrivons brièvement les propriétés de notre chaîne de traitement pré-syntaxique qui permet de gérer des corpus tout-venant. Nous présentons alors les deux systèmes d'analyse que nous avons utilisés, un analyseur TAG issu d'une méta-grammaire et un analyseur LFG. Nous comparons ces deux systèmes en indiquant leurs points communs, comme l'utilisation intensive du partage de calcul et des représentations compactes de l'information, mais également leurs différences, au niveau des formalismes, des grammaires et des analyseurs. Nous décrivons ensuite le processus de post-traitement, qui nous a permis d'extraire de nos analyses les informations demandées par la campagne EASy. Nous terminons par une évaluation quantitative de

nos architectures.

Direkt Profil est un analyseur automatique de textes écrits en français comme langue étrangère. Son but est de produire une évaluation du stade de langue des élèves sous la forme d'un profil d'apprenant. Direkt Profil réalise une analyse des phrases fondée sur des itinéraires d'acquisition, i.e. des phénomènes morpho-syntaxiques locaux liés à un développement dans l'apprentissage du français. L'article présente les corpus que nous traitons et d'une façon sommaire les itinéraires d'acquisition. Il décrit ensuite l'annotation que nous avons définie, le moteur d'analyse syntaxique et l'interface utilisateur. Nous concluons par les résultats obtenus jusqu'ici : sur le corpus de test, le système obtient un rappel de 83% et une précision de 83%.

Nous présentons un outil, ILIMP, qui prend en entrée un texte brut (sans annotation linguistique) rédigé en français et qui fournit en sortie le texte d'entrée où chaque occurrence du pronom il est décorée de la balise [ANAphorique] ou [IMPersonnel]. Cet outil a donc comme fonctionnalité de distinguer les occurrences anaphoriques du pronom il, pour lesquelles un système de résolution des anaphores doit chercher un antécédent, des occurrences où il est un pronom impersonnel (explétif) pour lequel la recherche d'antécédent ne fait pas sens. ILIMP donne un taux de précision de 97,5%. Nous présentons une analyse détaillée des erreurs et nous décrivons brièvement d'autres applications potentielles de la méthode utilisée dans ILIMP, ainsi que l'utilisation et le positionnement d'ILIMP dans un système d'analyse syntaxique modulaire.

Nous présentons ici une stratégie d'étiquetage et d'analyse syntaxique de que. Cette forme est en effet susceptible d'appartenir à trois catégories différentes et d'avoir de multiples emplois pour chacune de ces catégories. Notre objectif est aussi bien d'en assurer un étiquetage correct que d'annoter les relations de dépendance que que entretient avec les autres mots de la phrase. Les deux étapes de l'analyse mobilisent des ressources différentes.

Cet article s'intéresse au problème des erreurs orthographiques produisant des mots lexicalement corrects dans des textes en langue arabe. Après la description de l'influence des spécificités de la langue arabe sur l'augmentation du risque de commettre ces fautes cachées, nous proposons une classification hiérarchique de ces erreurs en deux grandes catégories ; à savoir syntaxique et sémantique. Nous présentons, également, l'architecture multi-agent que nous avons adoptée pour la détection et la correction des erreurs cachées en textes arabes. Nous examinons alors, les comportements sociaux des agents au sein de leurs organisations respectives et de leur environnement. Nous exposons vers la fin la mise en place et l'évaluation du système réalisé.

Cet article s'intéresse à la structure des représentations logiques des énoncés en langue naturelle. Par représentation logique, nous entendons une représentation sémantique incluant un traitement de la portée des quantificateurs. Nous montrerons qu'une telle représentation combine fondamentalement deux structures sous-jacentes, une structure « prédicative » et une structure hiérarchique logique, et que la distinction des deux permet, par exemple, un traitement élégant de la sous-spécification. Nous proposerons une grammaire polarisée pour manipuler directement la structure des représentations logiques (sans passer par un langage linéaire avec variables), ainsi qu'une grammaire pour l'interface sémantique-syntaxe.

Cet article dresse un aperçu du système MedSLT, un système de traduction de la parole dans le domaine médical pour un vocabulaire limité. Il met l'accent sur le problème du choix du type de représentation pour les constructions temporelles et causales. Nous montrons que celles-ci ne peuvent pas être représentées par des structures plates, généralement utilisées pour ce type d'application, mais qu'elles nécessitent des stuctures plus riches, enchâssées, qui permettent d'obtenir une traduction plus adéquate. Nous expliquons comment produire ces représentations et écrire des règles de traduction économiques qui mettent en correspondance les représentations

sources dans la représentation interlingue correspondante.

Nous montrons dans cet article qu'une même entité peut être désignée de multiples façons et que les noms désignant ces entités sont par nature polysémiques. L'analyse ne peut donc se limiter à une tentative de résolution de la référence mais doit mettre en évidence les possibilités de nommage s'appuyant essentiellement sur deux opérations de nature linguistique : la synecdoque et la métonymie. Nous présentons enfin une modélisation permettant de rendre explicite les différentes désignations en discours, en unifiant le mode de représentation des connaissances linguistiques et des connaissances sur le monde.

Nous décrivons un projet de production de résumé automatique de textes pour le domaine juridique pour lequel nous avons utilisé un corpus des jugements de la cour fédérale du Canada. Nous présentons notre système de résumé LetSum ainsi que l'évaluation des résumés produits. L'évaluation de 120 résumés par 12 avocats montre que la qualité des résumés produits par LetSum est comparable avec celle des résumés écrits par des humains.

Nous proposons une technique de résumé automatique de textes par contraction de phrases. Notre approche se fonde sur l'étude de la fonction syntaxique et de la position dans l'arbre syntaxique des constituants des phrases. Après avoir défini la notion de constituant, et son rôle dans l'apport d'information, nous analysons la perte de contenu et de cohérence discursive que la suppression de constituants engendre. Nous orientons notre méthode de contraction vers les textes narratifs. Nous sélectionnons les constituants à supprimer avec un système de règles utilisant les arbres et variables de l'analyse morpho-syntaxique de SYGFRAN [Cha84]. Nous obtenons des résultats satisfaisants au niveau de la phrase mais insuffisants pour un résumé complet. Nous expliquons alors l'utilité de notre système dans un processus plus général de résumé automatique.

Choi, Wiemer-Hastings & Moore (2001) ont proposé d'employer l'analyse sémantique latente (ASL) pour extraire des connaissances sémantiques à partir de corpus afin d'améliorer l'efficacité d'un algorithme de segmentation des textes. En comparant l'efficacité du même algorithme selon qu'il prend en compte des connaissances sémantiques complémentaires ou non, ils ont pu montrer les bénéfices apportés par ces connaissances. Dans leurs expériences cependant, les connaissances sémantiques avaient été extraites d'un corpus qui contenait les textes à segmenter dans la phase de test. Si cette hyperspécificité du corpus d'apprentissage explique la plus grande partie de l'avantage observé, on peut se demander s'il est possible d'employer l'ASL pour extraire des connaissances sémantiques génériques pouvant être employées pour segmenter de nouveaux textes. Les deux expériences présentées ici montrent que la présence dans le corpus d'apprentissage du matériel de test a un effet important, mais également que les connaissances sémantiques génériques dérivées de grands corpus améliorent l'efficacité de la segmentation.

Dans ce papier, nous présentons un système de Détection de Structures fines de Texte (appelé DST). DST utilise un modèle prédictif obtenu par un algorithme d'apprentissage qui, pour une configuration d'indices discursifs donnés, prédit le type de relation de dépendance existant entre deux énoncés. Trois types d'indices discursifs ont été considérés (des relations lexicales, des connecteurs et un parallélisme syntaxico-sémantique) ; leur repérage repose sur des heuristiques. Nous montrons que notre système se classe parmi les plus performants.

Les corpus parallèles sont d'une importance capitale pour les applications multi-lingues de traitement automatique des langues. Malheureusement, leur rareté est le maillon faible de plusieurs applications d'intérêt. Extraire de tels corpus duWeb est une solution viable, mais elle introduit une nouvelle problématique : il n'est pas toujours trivial d'identifier les documents parallèles parmi tous ceux qui ont été extraits. Dans cet article, nous nous intéressons à l'identification automatique des paires de documents parallèles contenues dans un corpus bilingue. Nous montrons que cette tâche

peut être accomplie avec précision en utilisant un ensemble restreint d'invariants lexicaux. Nous évaluons également notre approche sur une tâche de traduction automatique et montrons qu'elle obtient des résultats supérieurs à un système de référence faisant usage d'un lexique bilingue.

Cet article présente une méthode de traduction automatique statistique basée sur des segments non-continus, c'est-à-dire des segments formés de mots qui ne se présentent pas nécessairement de façon contiguë dans le texte. On propose une méthode pour produire de tels segments à partir de corpus alignés au niveau des mots. On présente également un modèle de traduction statistique capable de tenir compte de tels segments, de même qu'une méthode d'apprentissage des paramètres du modèle visant à maximiser l'exactitude des traductions produites, telle que mesurée avec la métrique NIST. Les traductions optimales sont produites par le biais d'une recherche en faisceau. On présente finalement des résultats expérimentaux, qui démontrent comment la méthode proposée permet une meilleure généralisation à partir des données d'entraînement.

Cet article présente et évalue une approche originale et efficace permettant d'aligner automatiquement un bi-texte au niveau des mots. Pour cela, cette approche tire parti d'une analyse syntaxique en dépendances des bi-textes effectuée par les outils SYNTEX et utilise une technique d'apprentissage artificiel, la programmation logique inductive, pour apprendre automatiquement des règles dites de propagation. Celles-ci se basent sur les informations syntaxiques connues pour ensuite aligner les mots avec une grande précision. La méthode est entièrement automatique, et les résultats évalués sur les données de la campagne d'alignement HLT montrent qu'elle se compare aux meilleures techniques existantes. De plus, alors que ces dernières nécessitent plusieurs millions de phrases pour s'entraîner, notre approche n'en requiert que quelques centaines. Enfin, l'examen des règles de propagation inférées permet d'identifier facilement les cas d'isomorphismes et de non-isomorphismes syntaxiques entre les deux langues traitées.

Cet article propose et évalue une méthode de traduction automatique de termes biomédicaux simples du français vers l'anglais et de l'anglais vers le français. Elle repose sur une technique d'apprentissage artificiel supervisée permettant d'inférer des transducteurs à partir d'exemples de couples de termes bilingues ; aucune autre ressource ou connaissance n'est requise. Ces transducteurs, capturant les grandes régularités de traduction existant dans le domaine biomédical, sont ensuite utilisés pour traduire de nouveaux termes français en anglais et vice versa. Les évaluations menées montrent que le taux de bonnes traductions de notre technique se situe entre 52 et 67%. À travers un examen des erreurs les plus courantes, nous identifions quelques limites inhérentes à notre approche et proposons quelques pistes pour les dépasser. Nous envisageons enfin plusieurs extensions à ce travail.

Lorsque nous écoutons un énoncé ou que nous lisons un texte, les phénomènes de saillance accaparent notre attention sur une entité du discours particulière. Cette notion de saillance comprend un grand nombre d'aspects, incluant des facteurs lexicaux, syntaxiques, sémantiques, pragmatiques, ou encore cognitifs. En tant que point de départ de l'interprétation du langage, la saillance fonctionne de pair avec la structure communicative. Dans cet article, notre but principal est de montrer comment aboutir à un modèle computationnel de la saillance, qui soit valable aussi bien pour la saillance linguistique que pour la saillance visuelle. Pour cela, nous retenons une liste de facteurs qui contribuent à rendre saillante une entité. Dans le cas d'une entité du discours, cette approche nous permet de clarifier les rapports entre saillance et structure communicative. Nous définissons nos principes de primordialité et de singularité, puis nous passons en revue les différentes méthodes de quantification de la saillance qui sont compatibles avec ces principes. Nous illustrons alors l'une d'entre elles sur un exemple linguistique et sur un exemple visuel.

Nous présentons dans cet article une extension de la SDRT (Segmented Discourse Representation Theory), pour un modèle d'interprétation pragmatique d'un système de dialogue homme-machine.

Partant d'une discussion sur les présupposés et les implicatures conversationnelles, nous analysons l'approche de Ducrot en Vue d'une intégration des topoï dans notre modèle. Nous y ajoutons la prise en compte des attentes dans le dialogue (effets projectifs des actes interlocutoires). Enfin nous proposons un mécanisme de résolution logique qui consiste à introduire plus systématiquement un noeud topique dans la SDRS (Discourse Represenlation Structure). Nous décrivons dans cet article les principes de traitement pragmatique mis en oeuvre, et nous illustrons le processus d'analyse à l'aide d'un exemple.

Ces dernières années, il y a eu de nombreux travaux portant sur l'utilisation d'actes de dialogue pour caractériser les dialogues homme-homme ou homme-machine. Cet article fait état de nos travaux sur la détection automatique d'actes de dialogue dans des corpus réels de dialogue homme-homme. Notre travail est fondé essentiellement sur deux hypothèses . (i) la position des mots et la classe sémantique du mot sont plus importants que les mots eux-mêmes pour identifier l'acte de dialogue et (ii) il y a une forte prédictivité dans la succession des actes de dialogues portés sur un même segment dialogique. Une approche de type Memory Based Learning a été utilisée pour la détection automatique des actes de dialogue. Le premier modèle n'utilise pas d'autres informations que celles contenus dans le tour de parole. Dans lex expériences suivantes, des historiques dialogiques de taille variables sont utilisés. Le taux d'erreur de détection d'actes de dialogue est d'environ 16% avec le premier modèle est descend avec une utilisation plus large de l'historique du dialogue à environ 14%.

Cet article propose une définition formelle de la notion de couverture lexicale. Celle-ci repose sur un ensemble de quatre métriques qui donnent une vue globale de l'adéquation d'une ressource lexicale à un corpus et permettent ainsi de guider le choix d'une ressource en fonction d'un corpus donné. Les métriques proposées sont testées dans le contexte de l'analyse de corpus spécialisés en génomique : 5 terminologies différentes sont confrontées à 4 corpus. La combinaison des valeurs

obtenues permet de discerner différents types de relations entre ressources et corpus.

Cette étude se place dans le cadre général de la désambiguïsation automatique du sens d'un Verbe dans un énoncé donné. Notre méthode de désambiguïsation prend en compte la construction du Verbe, c'est-à-dire l'influence des éléments lexicaux et syntaxiques présents dans l'énoncé (co-texte). Nous cherchons maintenant à finaliser cette méthode en tenant compte des caractéristiques sémantiques du co-texte. Pour ce faire nous associons au corpus un espace distributionnel continu dans lequel nous construisons et Visualisons des classes distributionnelles. La singularité de ces classes est qu'elles sont calculées à la Volée. Elles dépendent donc non seulement du corpus mais aussi du contexte étudié. Nous présentons ici notre méthode de calcul de classes ainsi que les premiers résultats obtenus.

L'identification et l'évaluation des avis, opinions ou jugements exprimés sur un sujet, une entreprise, ou un produit sont des tâches essentielles dans le domaine de l'analyse des médias. L'étude d'opinion est employée pour repérer de nouvelles tendances, mesurer le degré de satisfaction des clients ou pour alerter quand des tendances négatives risquent d'être défavorable à l'image de marque de l'entreprise. Dans cet article nous présentons un outil de veille économique qui permet de classifier très finement des documents publiés en ligne ainsi que d'identifier et d'évaluer les opinions exprimées dans des articles en ligne et des forums de discussions. Après la présentation des diverses composantes du système et des ressources linguistiques utilisées, nous décrivons en détail SentA, la composante d'étude d'opinions, et évaluons sa performance.

Nous proposons une réflexion théorique sur la place d'un phénomène tel que celui des disfluences au sein d'une grammaire. Les descriptions fines qui en ont été données mènent à se demander quel statut accorder aux disfluences dans une théorie linguistique complète, tout en conservant une perspective globale de représentation, c'est-à-dire sans nuire à la cohérence et à l'homogénéité

générale. Nous en introduisons une représentation formelle, à la suite de quoi nous proposons quelques mécanismes de parsing permettant de les traiter.

De nombreux linguistes ont mis en évidence des cas de « subordonnées » non dépendantes dans de multiples langues dans le monde (Mithun, 2003 ; Haiman & Thompson (eds), 1988). Ce phénomène a aussi été relevé en français, notamment pour un « subordonnant » tel que parce que (Debaisieux, 2001 ; Ducrot et al., 1975). Nous nous proposons de décrire un cas de « subordonnée » en quand non dépendante et de le représenter dans le cadre formel de Gerdes & Kahane (à paraître).

L'objectif de cet article est de montrer comment bâtir une structure de répresentation proche d'un graphe de dépendance à l'aide des deux structures de représentation canoniques fournies par les Grammaires d'Arbres Adjoints Lexicalisées . Pour illustrer cette approche, nous décrivons comment utiliser ces deux structures à partir d'une forêt partagée.

Cet article présente une évaluation de modèles statistiques du langage menée sur la langue Française. Nous avons cherché à comparer la performance de modèles de langage exotiques par rapport aux modèles plus classiques de n-gramme à horizon fixe. Les expériences réalisées montrent que des modèles de n-gramme à horizon variable peuvent faire baisser de plus de 10% en moyenne la perplexité d'un modèle de n-gramme à horizon fixe. Les modèles de n/m-multigramme demandent une adaptation pour pouvoir être concurrentiels.

La quasi-totalité des étiqueteurs grammaticaux mettent en oeuvre des règles qui portent sur les successions ou collocations permises de deux ou trois catégories grammaticales. Leurs performances s'établissent à hauteur de 96% de mots correctement étiquetés, et à moins de 57% de phrases correctement étiquetées. Ces règles binaires et tern-aires ne représentent qu'une

fraction du total des règles de succession que l'on peut extraire à partir des phrases d'un corpus d'apprentissage, alors même que la majeure partie des phrases (plus de 98% d'entre elles) ont une taille supérieure à 3 mots. Cela signifie que la plupart des phrases sont analysées au moyen de règles reconstituées ou simulées à partir de règles plus courtes, tern-aires en l'occurrence dans le meilleur des cas. Nous montrons que ces règles simulées sont majoritairement agrammaticales, et que l'avantage inférentiel qu'apporte le chaînage de règles courtes pour parer au manque d'apprentissage, plus marqué pour les règles plus longues, est largement neutralisé par la permissivité de ce processus dont toutes sortes de poids, scores ou probabilités ne réussissent pas à en hiérarchiser la production afin d'y distinguer le grammatical de l'agrammatical. Force est donc de reconsidérer les règles de taille supérieure à 3, lesquelles, il y a une trentaine d'années, avaient été d'emblée écartées pour des raisons essentiellement liées à la puissance des machines d'alors, et à l'insuffisance des corpus d'apprentissage. Mais si l'on admet qu'il faille désormais étendre la taille des règles de succession, la question se pose de savoir jusqu'à quelle limite, et pour quel bénéfice. Car l'on ne saurait non plus plaider pour une portée des règles aussi longue que les plus longues phrases auxquelles elles sont susceptibles d'être appliquées. Autrement dit, y a-t-il une taille optimale des règles qui soit suffisamment petite pour que leur apprentissage puisse converger, mais suffisamment longue pour que tout chaînage de telles règles pour embrasser les phrases de taille supérieure soit grammatical. La conséquence heureuse étant que poids, scores et probabilités ne seraient plus invoqués que pour choisir entre successions d'étiquettes toutes également grammaticales, et non pour éliminer en outre les successions agrammaticales. Cette taille semble exister. Nous montrons qu'au moyen d'algorithmes relativement simples l'on peut assez précisément la déterminer. Qu'elle se situe, compte tenu de nos corpus, aux alentours de 12 pour le français, de 10 pour l'arabe, et de 10 pour l'anglais. Qu'elle est donc en particulier inférieure à la taille moyenne des phrases, quelle que soit la langue considérée.

Cette étude est menée dans le cadre du développement de l'analyseur syntaxique de corpus Syntex

et porte sur la tâche de désambiguïsation des rattachements prépositionnels. Les données de sous-catégorisation syntaxique exploitées par Syntex pour la désambiguïsation se présentent sous la forme de probabilités de sous-catégorisation (que telle unité lexicale - verbe, nom ou adjectif - se construise avec telle préposition). Elles sont acquises automatiquement à partir d'un corpus de 200 millions de mots, étiqueté et partiellement analysé syntaxiquement. Pour évaluer ces données, nous utilisons 4 corpus de test de genres variés, sur lesquels nous avons annoté à la main plusieurs centaines de cas de rattachement prépositionnels ambigus. Nous testons plusieurs stratégies de désambiguïsation, une stratégie de base, une stratégie endogène qui exploite des propriétés de sous-catégorisation spécifiques acquises à partir du corpus en cours de traitement, une stratégie exogène qui exploite des propriétés de sous-catégorisation génériques acquises à partir du corpus de 200 millions de mots, et enfin une stratégie mixte qui utilisent les deux types de ressources. L'analyse des résultats montre que la stratégie mixte est la meilleure, et que les performances de l'analyseur sur la tâche de désambiguïsation des rattachements prépositionnels varient selon les corpus de 79.4 % à 87.2 %.

Cet article présente une méthode d'acquisition semi-automatique de collocations. Notre extraction monolingue estime pour chaque co-occurrence sa capacité à être une collocation, d'après une mesure statistique modélisant une caractéristique essentielle (le fait qu'une collocation se produit plus souvent que par hasard), effectue ensuite un filtrage automatique (en utilisant les vecteurs conceptuels) pour ne retenir que des collocations d'un certain type sémantique, puis effectue enfin un nouveau filtrage à partir de données entrées manuellement. Notre extraction bilingue est effectuée à partir de corpus comparables, et a pour but d'extraire des collocations qui ne soient pas forcément traductions mot à mot l'une de l'autre. Notre évaluation démontre l'intérêt de mêler extraction automatique et intervention manuelle pour acquérir des collocations et ainsi permettre de compléter les bases lexicales multilingues.

Cet article propose, au travers des résultats de différentes expériences sur la couverture des lexiques informatisés, de montrer que l'incomplétude lexicale est un phénomène constant dans tous les lexiques de TAL, mais que les mots inconnus eux-mêmes varient grandement selon les outils. Nous montrons également que la constance de cette incomplétude est étroitement liée à la créativité lexicale de la langue.

Nous proposons dans cet article une description de la Langue des Signes Française dans le but de traduire des énoncés courts du français et de les faire signer par un personnage de synthèse. Cette description pose en préalable la question de la transcription des éléments d'une langue dont le signal n'est pas linéaire. Il s'agit ensuite de repérer les différentes couches linguistiques et la forme de leurs unités constitutives en vue de la répartition des tâches informatiques : la synthèse de gestes nécessite un traitement des éléments constitutifs du geste et la génération syntaxique doit pouvoir manipuler des morphèmes.

L'extraction et la valorisation de données biographiques contenues dans les dépêches de presse est un processus complexe. Pour l'appréhender correctement, une définition complète, précise et fonctionnelle de cette information est nécessaire. Or, la difficulté que l'on rencontre lors de l'analyse préalable de la tâche d'extraction réside dans l'absence d'une telle définition. Nous proposons ici des conventions dans le but d'en développer une. Le principal concept utilisé pour son expression est la structuration de l'information sous forme de triplets {sujet, relation, objet}. Le début de définition ainsi construit est exploité lors de l'étape d'extraction d'informations par transducteurs à états finis. Il permet également de suggérer une solution d'implémentation pour l'organisation des données extraites en base de connaissances.

Dans cet article, nous cherchons à caractériser linguistiquement des segments textuels définis pragmatiquement, relativement à des besoins de réédition de documents et au sein desquels

l'information est susceptible d'évoluer dans le temps. Sur la base d'un corpus de textes encyclopédiques en français, nous analysons la distribution de marqueurs textuels et discursifs et leur pertinence en nous focalisant principalement sur un traitement sémantique particulier de la temporalité.

Nous présentons une méthode de traduction automatique de termes complexes pour la construction de ressources bilingues français/anglais, basée principalement sur une comparaison entre « mondes lexicaux » (ensemble de co-occurrents), à partir du Web. Nous construisons les mondes lexicaux des termes français sur le Web. Puis, nous générons leurs traductions candidates via un dictionnaire bilingue électronique et constituons les mondes lexicaux de toutes les traductions candidates. Nous comparons enfin les mondes lexicaux français et anglais afin de valider la traduction adéquate par filtres statistiques. Notre évaluation sur 10 mots français très polysémiques montre que l'exploitation des mondes lexicaux des termes complexes sur le Web permet une acquisition automatique de traductions avec une excellente précision.

Jusqu'à présent, la question de la reconnaissance automatique de métonymies a souvent été abordée avec des approches supervisées. Toutefois, ces approches nécessitent l'annotation d'un nombre important d'occurrences d'apprentissage et, dès lors, elles empêchent le développement d'un système de reconnaissance de métonymies à grande échelle. Cet article étudie la possibilité de résoudre ce problème du goulot d'étranglement de l'acquisition des connaissances en recourant à des techniques d'apprentissages non supervisées. Bien que la technique en question, l'algorithme de Schütze (1998), soit souvent appliquée en désambiguïsation sémantique, je montrerai qu'elle s'avère trop peu solide pour le cas spécifique de la reconnaissance de métonymies. À cet effet, je propose d'étudier l'influence de quatre variables sur les performances de la technique non supervisée, à savoir le type de données, la taille de la fenêtre d'observation, l'application de la décomposition en valeurs singulières (SVD) et le type de sélection de propriétés.

Cet article traite de l'utilité à détecter une chaîne coréférentielle de termes complexes afin d'améliorer la détection de variations de ce même terme complexe. Nous implémentons pour cela un programme permettant de détecter le nombre de variantes anaphoriques d'un terme complexe ainsi que le nombre de variantes anaphoriques de termes dans un texte scientifique. Ces deux fonctionnalités sont développées avec une ancrage dans une chaîne coréférentielle et en dehors de toute chaîne coréférentielle, afin de pouvoir évaluer l'efficacité de cette méthode.

Nous décrivons ici une approche pour passer d'une représentation syntaxique (issue d'une analyse grammaticale) à une représentation sémantique (sous forme de prédicats). Nous montrons ensuite que la construction de cette interface est automatisable. Nous nous appuyons sur l'interopérabilité de plusieurs ressources couvrant des aspects d'ordre syntaxique (Link Grammar Parser), lexical (WordNet) et syntaxico-sémantique (VerbNet) de la langue anglaise. L'utilisation conjointe de ces ressources de large couverture permet d'obtenir une désambiguïsation syntaxique et lexicale au moins partielle.

Le contexte est celui d'une plate-forme de génération automatique d'énoncés en langue signée, réalisés par un avatar 3D. Il existe quelques uns de ces systèmes aujourd'hui, par exemple le projet VisiCast (Hanke, 2002). Nous revenons ici sur les systèmes de description utilisés pour les unités gestuelles impliquées dans les énoncés, fondés sur un langage peu flexible et guère adaptatif. Nous proposons ensuite une nouvelle approche, constructiviste et géométrique, avec l'objectif de rendre la description des signes des lexiques signés plus adéquate, et par là améliorer leur intégration dans les discours générés.

L'interopérabilité entre les lexiques des grammaires d'unification passe par l'adoption d'une représentation normalisée de ces ressources. Dans ce papier, nous proposons l'utilisation de LMF

pour établir la standardisation des ressources lexicales en HPSG. Nous présentons LMF d'une manière sommaire et nous détaillons son utilisation pour coder les entrées lexicales d'un lexique HPSG.

Cet article décrit un système de construction du lexique et d'analyse morphologique pour l'arabe standard. Ce système profite des apports des modèles à états finis au sein de l'environnement linguistique de développement NooJ pour traiter aussi bien les textes voyellés que les textes partiellement ou non voyellés. Il se base sur une analyse morphologique faisant appel à des règles grammaticales à large couverture.

Le traitement linguistique d'un énoncé écrit conduit le plus souvent à la prise en compte d'interprétations concurrentes, ainsi qu'à la création d'ambiguïtés artificielles. Le contrôle de ces points d'embarras est indispensable pour garantir une efficacité et une précision convenable du processus d'analyse. L'approche décrite dans ce document exploite le paradigme de l'aide multicritère à la décision dans un contexte de TALN. Elle consiste à optimiser l'apport des méthodes spécifiques de contrôle dans une chaîne de traitement.

Nous présentons ici les bases d'une méthode de résolution de la coréférence entre les expressions nominales désignant des entités nommées. Nous comptons appliquer cet algorithme sur un corpus de textes journalistiques ; certains aspects de ce que l'on pourrait nommer les « facteurs de coréférence » dans ces textes nous amènent à favoriser l'utilisation de méthodes statistiques pour accomplir cette tâche. Nous décrivons l'algorithme de résolution de la coréférence mis en oeuvre, constitué d'un classifieur bayésien naïf.

L'apprentissage automatique de la sémantique est un sujet assez populaire dans le domaine du traitement automatique du langage. Beaucoup de recherches ont été éffectuées en comparant des

contextes syntaxiques similaires. On peut, par exemple, trouver des substantifs d'un champ sémantique similaire en examinant les adjectifs avec lesquels ils sont souvent en relation. Si on opte pour cette méthode, il y a néanmoins deux problèmes qui se posent, à savoir la complexité computationnelle et l'insuffisance des données. Cet article décrit l'application d'une technique mathématique, la décomposition en valeurs singulières. Cette technique a été appliquée au domaine de Recherche d'Information avec des résultats favorables. On se demande s'il est possible de trouver, grâce à la technique, des dimensions sémantiques latentes à l'espace d'adjectifs reduit avec lesquelles on peut faire un groupement qui est aussi bon ou meilleur que le groupement original.

Nous présentons dans cet article un analyseur sémantique pour la langue arabe. Cet analyseur contribue à la sélection du sens adéquat parmi l'ensemble des sens possibles que peut recevoir un mot hors contexte. Pour atteindre cet objectif, nous proposons un modèle vectoriel qui permet de lever les ambiguïtés locales au niveau de la phrase et celles relevant du domaine. Ce modèle est inspiré des modèles vectoriels très utilisés dans le domaine de la recherche documentaire.

De nouveaux outils de correction linguistique sont disponibles pour le français depuis quelques mois. Mis à la disposition des utilisateurs de Microsoft Office 2003, un nouveau correcteur orthographique et un nouveau correcteur grammatical permettent d'améliorer le processus de rédaction de documents. En partant d'évaluations externes effectuées récemment, nous présentons les diverses facettes de ces améliorations et de ces outils, en abordant la question de l'évaluation des outils de correction linguistique (qu'évaluer ? quels critères appliquer ? pourquoi développer une nouvelle version ?). La réforme de l'orthographe, la féminisation des noms de métier, l'évolution de la langue figurent parmi les thèmes abordés dans cet article.

Les systèmes de traitement du langage naturel qui combinent des méthodes basées sur les règles

(connaissances explicitées) et celles basées sur les corpus deviennent suffisamment précis pour permettre leur utilisation dans des applications variées. Nous décrivons un système d'analyse syntaxique de ce type pour le néerlandais, appelé Alpino, et nous montrons la nécessité d'utiliser des méthodes basées sur l'utilisation des corpus en vue d'obtenir des analyseurs par règles fiables. Nous décrivons plus particulièrement un ensemble de cas où les résultats de l'analyseur sont exploités pour améliorer l'analyseur lui-même.

Nous prouvons que l'information mutuelle entre des paires de mots peut être employée avec succès pour distinguer entre différents usages des mots dans la traduction des requêtes pour la recherche d'information translinguistique. Les expérmentations sont entreprises dans le contexte de la recherche d'information translinguistique amhariquefrancais. Des expérmentations sont entreprises qui comparent la performance de la collection des termes des requêtes désambiguïsés et non désambiguïsés contre une collection de documents ordonnés. Les résultats montrent une amélioration de performance pour les requêtes désambiguïsées en comparison avec l'approche alternative qui emploie la collection de termes entièrement expansés.

La désambiguïsation lexicale a une place centrale dans les applications de Traitement Automatique des Langues relatives à la traduction. Le travail présenté ici fait partie d'une étude sur les recouvrements et les divergences entre les espaces sémantiques occupés par des unités polysémiques de deux langues. Les correspondances entre ces unités sont rarement biunivoques et l'étude de ces correspondances aide à tirer des conclusions sur les possibilités et les limites d'utilisation d'une autre langue pour la désambiguïsation des unités d'une langue source. Le but de ce travail est l'établissement de correspondances d'une granularité optimale entre les unités de deux langues entretenant des relations de traduction. Ces correspondances seraient utilisables pour la prédiction des équivalents de traduction les plus adéquats de nouvelles occurrences des éléments polysémiques.

Dans cet article nous présentons un analyseur morphologique pour le verbe akkadien. Cette langue est de la famille des langues sémitiques. Les flexions du verbe font intervenir des changements internes à la racine. L'analyseur présenté ici illustre l'utilisation d'un formalisme multi-niveaux et d'opérateurs relationnels puissants, notamment la jointure. La multiplicité de niveaux intermédiaires entre les formes profondes et de surface, ainsi que les opérateurs de compositions permettent de diviser la description en contraintes relativement simples qui sont ensuite rassemblées pour s'exercer soit simultanément, soit en cascade, soit encore d'une façon mixte, c'est-à-dire simultanément pour certains des niveaux et en cascade pour d'autres. Ce mécanisme nous permet de décrire la vocalisation du radical comme un processus d'insertions successives de voyelles. Cela présente l'intérêt d'être plus simple que l'utilisation d'un schéma vocalique figé soumis à interdigitation. De plus, cela semble expliquer de façon plus économique les formes des verbes faibles.

Cet article présente la méthodologie et les résultats d'une analyse sémantique quantitative d'environ 5000 spécificités dans le domaine technique des machines-outils pour l'usinage des métaux. Les spécificités seront identifiées avec la méthode des mots-clés (KeyWords Method). Ensuite, elles seront soumises à une analyse sémantique quantitative, à partir du recouvrement des co-occurrences des co-occurrences, permettant de déterminer le degré de monosémie des spécificités. Finalement, les données quantitatives de spécificité et de monosémie feront l'objet d'analyses de régression. Nous avançons l'hypothèse que les mots (les plus) spécifiques du corpus technique ne sont pas (les plus) monosémiques. Nous présenterons ici les résultats statistiques, ainsi qu'une interprétation linguistique. Le but de cette étude est donc de vérifier si et dans quelle mesure les spécificités du corpus technique sont monosémiques ou polysémiques et quels sont les facteurs déterminants.

La plate-forme logicielle Outilex, qui sera mise à la disposition de la recherche, du développement et de l'industrie, comporte des composants logiciels qui effectuent toutes les opérations fondamentales du traitement automatique du texte écrit : traitements sans lexiques, exploitation de lexiques et de grammaires, gestion de ressources linguistiques. Les données manipulées sont structurées dans des formats XML, et également dans d'autres formats plus compacts, soit lisibles soit binaires, lorsque cela est nécessaire ; les convertisseurs de formats nécessaires sont inclus dans la plate-forme ; les formats de grammaires permettent de combiner des méthodes statistiques avec des méthodes fondées sur des ressources linguistiques. Enfin, des lexiques du français et de l'anglais issus du LADL, construits manuellement et d'une couverture substantielle seront distribués avec la plate-forme sous licence LGPL-LR.

Aujourd'hui, l'approche la plus courante en traitement de la parole consiste à combiner un reconnaisseur statistique avec un analyseur robuste. Pour beaucoup d'applications cependant, les reconnaisseurs linguistiques basés sur les grammaires offrent de nombreux avantages. Dans cet article, nous présentons une méthodologie et un ensemble de logiciels libres (appelé Regulus) pour dériver rapidement des reconnaisseurs linguistiquement motivés à partir d'une grammaire générale partagée pour le catalan et le français.

Nous présentons dans cette étude un essai de prise en compte des disfluences dans un système d'analyse linguistique initialement prévu pour l'écrit, en vue de la réalisation d'un prototype de traduction parole-parole. À partir d'une étude approfondie sur corpus, nous montrons comment des modifications du lexique et de la grammaire ont permis de traiter les cas les plus simples (pauses remplies, répétitions de mots isolés, etc.). D'autres cas plus complexes comme répétitions et auto-corrections de syntagmes ont nécessité la mise au point d'un mécanisme de contrôle sémantique permettant de limiter la combinatoire. Cette étude a mis également en évidence la difficulté de traitement de phénomènes tels que les amorces (mots interrompus) et les constructions

inachevées, qui pour l'instant restent sans solution satisfaisante.

Nous montrons une utilisation du Web, corpus multi-lingue de grande taille, pour effectuer une acquisition supervisée de concepts bilingue français/anglais. Cette acquisition utilise comme point initial un verbe français. Nous apparions ensuite des phrases provenant des deux langues à partir de couples de noms propres possédant la même forme dans les deux langues. Cet appariement automatique mais sommaire ne garantit pas l'alignement des phrases. Nous montrons qu'il nous permet cependant d'extraire des termes français et anglais équivalents dans leur contexte d'utilisation. Ces termes constituent des ressources multi-lingues particulièrement adaptées au Web, notamment pour les applications question réponse « cross-lingue ».

Nous étudions les relations de proximité sémantique entre les noms et les verbes à partir de données calculées sur un corpus de 200 millions de mots par un programme d'analyse distributionnelle automatique. Nous exposons les résultats d'une méthode d'extraction de couples Nom/Verbe, qui combine un indice de proximité distributionnelle et un indice de co-occurrence : un couple est extrait si le nom et le verbe apparaissent avec les mêmes arguments sur l'ensemble du corpus, d'une part, et s'ils apparaissent au moins une fois dans un même paragraphe munis du même argument, d'autre part. L'article élabore une typologie des 1441 couples extraits et démontre l'intérêt de prendre en compte les couples non liés morphologiquement, qui constituent 70 % des données.

Nous présentons un système de synthèse d'information pour la production de résumés multi-documents orientés par une requête complexe. Après une analyse du profil de l'utilisateur exprimé par des questions complexes, nous comparons la similarité entre les documents à résumer avec les questions à deux niveaux : global et détaillé. Cette étude démontre l'importance d'étudier pour une requête la pertinence d'une phrase à l'intérieur de la structure thématique du document.

Cette méthodologie a été appliquée lors de notre participation à la campagne d'évaluation DUC 2005 où notre système a été classé parmi les meilleurs.

Les tables du LADL (Laboratoire d'Automatique Documentaire et Linguistique) contiennent des données électroniques extensives sur les propriétés morpho-syntaxiques et syntaxiques des foncteurs syntaxiques du français (verbes, noms, adjectifs). Ces données, dont on sait qu'elles sont nécessaires pour le bon fonctionnement des systèmes de traitement automatique des langues, ne sont cependant que peu utilisées par les systèmes actuels. Dans cet article, nous identifions les raisons de cette lacune et nous proposons une méthode de conversion des tables vers un format mieux approprié au traitement automatique des langues.

Dans cet article, nous considérons un formalisme linguistique pour lequel l'intégration d'information sémantique dans une grammaire à large couverture n'a pas encore été réalisée à savoir, les grammaires d'arbres adjoints (Tree Adjoining Grammar ou TAG). Nous proposons une méthode permettant cette intégration et décrivons sa mise en oeuvre dans une grammaire noyau pour le français. Nous montrons en particulier que le formalisme de spécification utilisé, XMG, (Duchier et al., 2004) permet une factorisation importante des données sémantiques facilitant ainsi le développement, la maintenance et le déboggage de la grammaire.

Dans cet article, nous présentons une approche afin de traiter les questions booléennes, c'est-à-dire des questions dont la réponse peut être un Oui ou un Non, cela, dans le cadre d'un système de Questions-Réponses. En effet, la campagne Technolangue-EQueR, première campagne francophone de Questions-Réponses (QR) utilisant des questions et un corpus en français, a également été la première campagne QR à introduire une évaluation pour ce type de questions. Nous détaillons, parallèlement à notre approche, des pistes de réflexion sur les aspects sous-jacents à ces questions booléennes, notamment au travers d'une analyse des résultats

obtenus par notre système dans un contexte similaire à celui de notre participation à la campagne officielle.

Dans ce travail, nous étudions en corpus la productivité quantitative des suffixations par -Able et par -ité du français, d'abord indépendamment l'une de l'autre, puis lorsqu'elles s'enchaînent dérivationnellement (la suffixation en -ité s'applique à des bases en -Able dans environ 15 % des cas). Nous estimons la productivité de ces suffixations au moyen de mesures statistiques dont nous suivons l'évolution par rapport à la taille du corpus. Ces deux suffixations sont productives en français moderne : elles forment de nouveaux lexèmes tout au long des corpus étudiés sans qu'on n'observe de saturation, leurs indices de productivité montrent une évolution stable bien qu'étant dépendante des calculs qui leur sont appliqués. On note cependant que, de façon générale, de ces deux suffixations, c'est la suffixation par -ité qui est la plus fréquente en corpus journalistique, sauf précisément quand -ité s'applique à un adjectif en -Able. Étant entendu qu'un adjectif en -Able et le nom en -ité correspondant expriment la même propriété, ce résultat indique que la complexité de la base est un paramètre à prendre en considération dans la formation du lexique possible.

Nous proposons de traiter la coordination comme un entassement paradigmatique, établissant une relation de parataxe entre ses constituants. Par cette considération et ses implications sur la description et l'analyse, on s'éloigne des assomptions les plus fréquentes en linguistique formelle sur le traitement de la coordination. Nous introduisons une description des caractéristiques syntaxiques de cette proposition, ainsi que sa représentation formelle et son intégration au sein d'une grammaire du français qui a pour objet d'être utilisée en traitement automatique. Cette description strictement syntaxique a vocation à être complétée par des informations provenant d'autres domaines, ce qui nous permet d'illustrer quelques spécificités notables de notre modèle.

Pour améliorer l'efficacité des systèmes de recherche d'informations précises, l'utilisation de

connaissances sémantiques est nécessaire. Cependant pour le français, les outils de connaissances sémantiques telles les thesaurus sur domaine ouvert ne sont d'une part pas très nombreux et d'autre part pas suffisamment complets. Dans cet article, nous expliquons premièrement, l'intérêt de l'utilisation de connaissances sémantiques pour un système de question réponse. Puis, nous présentons le thesaurus EuroWord-Net, notamment ses limites et les améliorations que nous avons effectuées pour la base française dans un souci de le rendre plus satisfaisant pour notre application par l'ajout de relations inexistantes entre concepts et de définitions par le biais de l'encyclopédieWikipedia (2006).

Notre objectif est la reconnaissance automatique de certaines formes dérivées, i.e. des diminutifs et des augmentatifs des noms et des adjectifs simples, ainsi que des comparatifs et des superlatifs des adjectifs simples du grec moderne. Il s'agit de formes qui sont généralement produites par l'adjonction d'un suffixe à la forme standard correspondante. Nous justifions notre choix de les ajouter dans le dictionnaire électronique. Leur traitement a nécessité une nouvelle représentation du dictionnaire qui utilise désormais un système de règles permettant de générer aisément les formes fléchies dérivées, de les étiqueter en tant que telles, et de les mettre en relation avec leur forme de base. Il en résulte une meilleure structuration des ressources lexicales et une production de dictionnaires flexible.

La résolution des anaphores dans les systèmes de dialogue homme-machine s'inspire généralement des modèles et des algorithmes développés pour le texte. Or le dialogue met en jeu une situation, c'est-à-dire un environnement physique immédiat et des événements dont la perception est partagée par les interlocuteurs. Cette situation peut servir d'ancrage à des expressions référentielles dites « anaphores à antécédents non linguistiques ». L'attribution de référents à de telles expressions s'avère difficile pour deux raisons : premièrement les facteurs situationnels sont nombreux et peu explicites ; deuxièmement des ambiguïtés peuvent apparaître

entre de possibles antécédents situationnels et de possibles antécédents linguistiques. Nous proposons ici un modèle clarifiant l'intervention des facteurs situationnels et permettant leur prise en compte lors de la compréhension des expressions référentielles potentiellement anaphoriques. En intégrant la notion de saillance valable à la fois pour les aspects situationnels et linguistiques, nous montrons comment utiliser des scores numériques pour gérer les interférences entre hypothèses situationnelles et linguistiques.

Nous présentons notre participation à la deuxième campagne d'évaluation de CESTA, un projet EVALDA de l'action Technolangue. Le but de cette campagne consistait à tester l'aptitude des systèmes de traduction à s'adapter rapidement à une tâche spécifique. Nous analysons la fragilité d'un système de traduction probabiliste entraîné sur un corpus hors-domaine et dressons la liste des expériences que nous avons réalisées pour adapter notre système au domaine médical.

Les systèmes de question-réponse sont la plupart du temps composés de trois grands modules : l'analyse de la question, la sélection des documents et l'extraction de la réponse. Dans cet article, nous nous intéressons au troisième module, plus particulièrement dans le cas plus délicat où la réponse attendue n'est pas du type entitée nommée. Nous décrivons comment l'analyseur Cass est employé pour marquer la réponse dans les phrases candidates et nous évaluons les résultats de cette approche. Au préalable, nous décrivons et évaluons le module dédié à l'analyse de la question, car les informations qui en sont issues sont nécessaires à notre étape finale d'extraction.

Dans le traitement automatique du langage naturel, les dictionnaires électroniques associent à chaque mot de l'information. La représentation informatique la plus efficace de ces dictionnaires utilise des machines à nombre fini d'états (automates ou transducteurs). Dans cet article, nous nous inspirons des algorithmes de construction directe d'un automate déterministe minimal pour proposer une nouvelle forme de transducteur. Cette nouvelle forme permet un calcul rapide des sorties

associées aux mots, tout en étant plus compacte quant au nombre de transitions et de sorties distinctes, comme le montrent nos expérimentations.

Nous développons un système qui doit être capable d'effectuer les mêmes inférences que le lecteur humain d'un constat d'accident de la route, et plus particulièrement de déterminer les causes apparentes de l'accident. Nous décrivons les niveaux linguistiques et sémantiques de l'analyse, et les règles d'inférence utilisées par ce système.

Cet article étudie la résolution des références à des entités lorsqu'une représentation informatique de ces entités est disponible. Nous nous intéressons à un corpus de dialogues entre humains, portant sur les grands titres de la presse francophone du jour, et proposons une méthode pour détecter et résoudre les références faites par les locuteurs aux articles des journaux. La détection des expressions nominales qui réfèrent à ces documents est réalisée grâce à une grammaire, alors que le problème de la détection des pronoms qui réfèrent aux documents est abordé par des moyens statistiques. La résolution de ces expressions, à savoir l'attribution des référents, fait quant à elle l'objet d'un algorithme inspiré de la résolution des co-références. Ces propositions sont évaluées par le biais de mesures quantitatives spécifiques.

L'acquisition automatique sur corpus d'informations lexicales sémantiques donne une place importante à la constitution de classes sémantiques rassemblant des mots de sens proches. Or, l'intérêt pratique de celles-ci reste limité en l'absence d'information sur les distinctions individualisant les sens des mots qu'elles rassemblent. Nous présentons dans cet article un premier système permettant de mettre au jour, de manière semi-automatique et à partir des seules données textuelles rassemblées dans un corpus, des éléments de distinction sémantique fine entre mots appartenant à une même classe, atteignant ainsi un degré de définition du sens encore inédit en acquisition automatique d'informations sémantiques lexicales. La technique mise au point regroupe,

en s'appuyant sur l'étude de grands voisinages autour des occurrences des mots comparés, des paires de mots distingués par des nuances similaires. Cette approche présente la faiblesse de ne permettre qu'une représentation implicite des nuances découvertes : les listes de paires de mots rapprochées doivent être interprétées afin de « comprendre » l'élément de distinction commun. En revanche, elle permet une automatisation importante du processus de recherche de nuances, suffisante pour assurer que le travail humain de validation des résultats n'introduise dans ceux-ci de biais interprétatif trop important.

L'identification des structures prédicatives présente un grand intérêt quand on se situe dans une problématique d'extraction d'information. Si une littérature abondante existe à ce sujet, particulièrement dans le domaine de la génomique, la plupart des travaux portent sur les relations autour du verbe. Peu s'intéressent à la relation qui peut unir une nominalisation et ses actants dans un groupe nominal à tête prédicative (GNP). Nous montrons la complexité des différents types de GNP et des relations paraphrastiques qui les unissent avec les formes verbales, afin de donner une vue unifiée des structures prédicatives nomino-verbales. Nous montrons ensuite comment nous avons conçu une grammaire de liens permettant l'identification de chacun des actants dans les GNP. Nous en décrivons la mise en oeuvre avec le Link Parser, pour l'extraction d'information dans des articles scientifiques du domaine de la Biologie.

Nous présentons une méthode de fouille d'erreurs pour détecter automatiquement des erreurs dans les ressources utilisées par les systèmes d'analyse syntaxique. Nous avons mis en oeuvre cette méthode sur le résultat de l'analyse de plusieurs millions de mots par deux systèmes d'analyse différents qui ont toutefois en commun le lexique syntaxique et la chaîne de traitement pré-syntaxique. Nous avons pu identifier ainsi des inexactitudes et des incomplétudes dans les ressources utilisées. En particulier, la comparaison des résultats obtenus sur les sorties des deux analyseurs sur un même corpus nous a permis d'isoler les problèmes issus des ressources

partagées de ceux issus des grammaires.

Partant des lexiques TAL syntaxiques existants, cet article propose une représentation lexicale unifiée et normalisée, préalable et nécessaire à toute exploitation des lexiques syntaxiques hors de leur propre contexte de conception. Ce travail s'inscrit dans un cadre de modélisation privilégié ? le Lexical Markup Framework ? qui a été conçu dès le départ comme un modèle lexicographique intégrant les différents niveaux de description. Ce modèle permet d'articuler des descriptions extensionnelles et intensionnelles et fait référence à un jeu de descripteurs normalisés, garantissant la rigueur de la description des faits linguistiques et assurant, à terme, la compatibilité avec des formats de données utilisés pour l'annotation de corpus.

Dans cet article nous présentons un modèle déductif-inductif pour l'identification des typologies textuelles et des genres dans les pages Web. Dans ce modèle, les typologies textuelles sont déduites en utilisant une forme modifiée du théorème de Bayes, tandis que les genres sont dérivés au moyen de simples règles « si-alors ». Étant donné que le système des genres sur le Web est complexe et que les pages Web sont plus imprévisibles et individualisées que les documents traditionnels, nous proposons cette approche déductive-inductive comme une alternative aux méthodes statistiques supervisées et non-supervisées. En effet, le modèle déductif-inductif permet une classification qui peut s'accommoder des genres non complètement standardisés. Il est aussi plus respectueux à l'égard de la vraie nature de la page Web, qui est en fait mixte et ne correspond presque jamais à un type idéal ou à un prototype précis, mais présente plutôt un mélange de genres, ou pas de genre du tout. L'évaluation de ce modèle reste un problème à résoudre.

La recherche d'information consiste à trouver les documents pertinents parmi un ensemble de documents en réponse à une requête de l'utilisateur. Ces documents sont triés par ordre de pertinence. Le but du traitement automatique du langage naturel dans la recherche d'information est

de transformer les mots potentiellement ambigus de la requête et des documents en représentations internes non ambiguës sur lesquelles s'effectuera l'appariement. Cette transformation est généralement réalisée à l'aide de plusieurs niveaux d'analyse linguistique (morphologique, syntaxique, etc.). Cet article présente l'analyseur linguistique de l'arabe du moteur de recherche cross-lingue du LIC2M. Nous allons nous concentrer sur l'analyseur morphologique et plus particulièrement sur le module de segmentation qui permet de découper les mots agglutinés en proclitiques, formes simples et enclitiques. Nous allons démontrer qu'une bonne segmentation améliore la précision et le rappel du moteur de recherche.

Cette étude présente les travaux du LIA effectués sur le corpus de dialogue homme-machine MEDIA et visant à proposer des méthodes d'analyse robuste permettant d'extraire d'un message audio une séquence de concepts élémentaires. Le modèle de décodage conceptuel présenté est basé sur une approche stochastique qui intègre directement le processus de compréhension au processus de Reconnaissance Automatique de la Parole (RAP). Cette approche permet de garder l'espace probabiliste des phrases produit en sortie du module de RAP et de le projeter vers un espace probabiliste de séquences de concepts. Les expériences menées sur le corpus MEDIA montrent que les performances atteintes par notre modèle sont au niveau des meilleurs systèmes ayant participé à l'évaluation sur des transcriptions manuelles de dialogues. En détaillant les performances du système en fonction de la taille du corpus d'apprentissage on peut mesurer le nombre minimal ainsi que le nombre optimal de dialogues nécessaires à l'apprentissage des modèles. Enfin nous montrons comment des connaissances a priori peuvent être intégrées dans nos modèles afin d'augmenter significativement leur couverture en diminuant, à performance égale, l'effort de constitution et d'annotation du corpus d'apprentissage.

Nous proposons et testons deux méthodes de prédiction de la capacité d'un système à répondre à une question factuelle. Une telle prédiciton permet de déterminer si l'on doit initier un dialogue afin

de préciser ou de reformuler la question posée par l'utilisateur. La première approche que nous proposons est une adaptation d'une méthode de prédiction dans le domaine de la recherche documentaire, basée soit sur des machines à vecteurs supports (SVM) soit sur des arbres de décision, avec des critères tels que le contenu des questions ou des documents, et des mesures de cohésion entre les documents ou passages de documents d'où sont extraits les réponses. L'autre approche vise à utiliser le type de réponse attendue pour décider de la capacité du système à répondre. Les deux approches ont été testées sur les données de la campagne Technolangue EQUER des systèmes de questions-réponses en français. L'approche à base de SVM est celle qui obtient les meilleurs résultats. Elle permet de distinguer au mieux les questions faciles, celles auxquelles notre système apporte une bonne réponse, des questions difficiles, celles restées sans réponses ou auxquelles le système a répondu de manière incorrecte. A l'opposé on montre que pour notre système, le type de réponse attendue (personnes, quantités, lieux...) n'est pas un facteur déterminant pour la difficulté d'une question.

Cet article présente un travail destiné à automatiser l'étude de l'évolution terminologique à partir de termes datés extraits de corpus diachroniques de textes scientifiques ou techniques. Les apparitions et disparitions d'attestations de termes au cours du temps constituent la manifestation la plus simple de l'évolution. Mais la prise en compte des formes variantes apporte une information de meilleure qualité sur le suivi des termes. Une distance entre termes complexes permet de rendre opérationnelle l'intégration de la variation terminologique à l'analyse diachronique. Des résultats montrant la prise en compte des variantes sont présentés et commentés à la fin de l'article.

Ce papier expose une expérience de classification menée sur un corpus de définitions dictionnairiques. Le cadre général de cette recherche est la constitution d'une ressource lexico-sémantique fondée sur une conception structuraliste du sens (le contenu sémantique d'une unité lexicale est structuré en sèmes ; le sens d'un texte émerge de faisceaux de regroupements

sémiques stabilisés). L'objectif de l'expérience rapportée est de découvrir des classes sémantiques à partir de définitions dictionnairiques avec la méthode CAH. Les classes sémantiques regroupent des unités lexicales en fonction de sèmes génériques (i.e. communs à toutes les unités lexicales de la classe) et s'organisent différentiellement en fonction de sèmes spécifiques. À partir d'une sélection d'entrées dictionnairiques partageant le sème générique /arbre/, nous étudions la distribution et l'organisation d'une hypothétique classe sémantique liée au domaine de la sylviculture.

Nous abordons ici la question de l'analyse de la structure du discours, du point de vue de sa description formelle et de son traitement automatique. Nous envisageons l'hypothèse selon laquelle une approche par contraintes pourrait permettre la prise en charge de structures discursives variées d'une part, et de différents types d'indices de leur manifestation d'autre part. Le formalisme CDML que nous introduisons vise précisément une telle approche.

Dans cet article, nous proposons une démarche d'analyse syntaxique pour les phrases nominales arabes à l'aide du formalisme des grammaires syntagmatiques guidées par les têtes HPSG. Pour ce faire, nous commençons par étudier la typologie de la phrase nominale arabe en précisant ses différentes formes. Puis, nous élaborons une grammaire HPSG traitant ce type de phrase et qui respecte la spécificité de la langue arabe. Ensuite, nous présentons une démarche d'analyse syntaxique se basant sur une approche ascendante et sur le mécanisme d'unification. Enfin, nous donnons une idée sur l'implémentation et l'expérimentation du prototype réalisé.

Cet article présente la méthodologie suivie et les résultats obtenus dans le cadre d'un projet qui a pour objectif la construction d'une large base de données d'expressions multi-mots de la langue portugaise. Ces expressions multi-mots ont été automatiquement extraites d'un corpus équilibré de 50 millions de mots, interprétées statistiquement à l'aide de mesures d'association lexicales et ont

été ensuite manuellement vérifiées. La base de données lexicales recouvre différent types d'expressions multi-mots avec différents degrés de cohésion, qui vont de la quasi totale fixité jusqu'aux groupes de mots qui se réalisent préférentiellement ensemble, comme les collocations. Le large ensemble de données de cette ressource permettra une révision des typologies d'unités multi-mots en portugais et l'évaluation de différentes mesures d'associations lexicales.

L'interface homme-machine a besoin de modèles de structures de dialogue qui expliquent la variabilité et la spontanéité au dialogue. Le contexte sémantique et pragmatique évolue continuellement pendant le développement de la conversation, surtout par la distribution de turns qui ont un effet direct dans les échanges de dialogue. Dans cet article nous utilisons un paradigme de langue formel pour modéliser les conversations de système de multiagent. Notre modèle computationnel combine des unités minimales pragmatiques -les actes de parole- pour construire des dialogues. Dans ce cadre, nous montrons comment la distribution de turn-taking peut être ambiguë et proposer un algorithme pour la résoudre, considérant turn coherence, trajectories et le turn-pairing. Finalement, nous suggérons overlapping comme un des phénomènes possibles naissants d'un turn-taking non résolu.

L'objectif de cette recherche est d'évaluer l'efficacité d'algorithmes lors de l'identification des ruptures thématiques dans des textes. Pour ce faire, 32 articles de journaux ont été segmentés par des groupes de 15 juges. L'analyse de leurs réponses indique que chaque juge, pris individuellement, est peu fiable contrairement à l'indice global de segmentation, qui peut être dérivé des réponses de l'ensemble des juges. Si les deux algorithmes testés sont capables de retrouver le début des articles lorsque ceux-ci sont concaténés, ils échouent dans la détection des changements de thème perçus par la majorité des juges. Il faut toutefois noter que les juges, pris individuellement, sont eux-mêmes inefficaces dans l'identification des changements de thème. Dans la conclusion, nous évaluons différentes explications du faible niveau de performance observé.

Les méthodes d'analyse syntaxiques hybrides, reposant à la fois sur des techniques statistiques et symboliques, restent peu exploitées. Dans la plupart des cas, les informations statistiques sont intégrées à un squelette context-free et sont utilisées pour contrôler le choix des règles ou des structures. Nous proposons dans cet article une méthode permettant de calculer un indice de corrélation entre deux objets linguistiques (catégories, propriétés). Nous décrivons une utilisation de cette notion dans le cadre de l'analyse des Grammaires de Propriétés. L'indice de corrélation nous permet dans ce cas de contrôler à la fois la sélection des constituants d'une catégorie, mais également la satisfaction des propriétés qui la décrivent.

Dans le cadre de la modélisation statistique du langage, nous montrons qu'il est possible d'utiliser un modèle n-grammes avec un historique qui n'est pas nécessairement celui avec lequel il a été appris. Par exemple, un adverbe présent dans l'historique peut ne pas avoir d'importance pour la prédiction, et devrait donc être ignoré en décalant l'historique utilisé pour la prédiction. Notre étude porte sur les modèles n-grammes classiques et les modèles n-grammes distants et est appliquée au cas des bi-grammes. Nous présentons quatre cas d'utilisation pour deux modèles bi-grammes : distants et non distants. Nous montrons que la combinaison linéaire dépendante de l'historique de ces quatre cas permet d'améliorer de 14 % la perplexité du modèle bi-grammes classique. Par ailleurs, nous nous intéressons à quelques cas de combinaison qui permettent de mettre en valeur les historiques pour lesquels les modèles que nous proposons sont performants.

Cet article présente l'utilisation de « Jibiki » (la plateforme de développement du serveur Web Papillon) dans le cadre du projet LexALP 1. Le but de ce projet est d'harmoniser la terminologie des quatre langues (français, allemand, italien et slovène) de la Convention Alpine 2 de sorte que les états membres puissent coopérer efficacement. Pour cela, le projet utilise la plateforme Jibiki afin de construire une banque terminologique permettant de comparer la terminologie spécialisée de sept

systèmes légaux dans quatre langues, et de l'harmoniser, optimisant ainsi la compréhension entre les états alpins sur des questions environnementales au niveau supranational. Dans cet article, nous présentons comment peut être employée la plateforme générique Jibiki afin de gérer un dictionnaire particulier.

Traiter les erreurs en dialogue homme-machine est un problème difficile compte-tenu des multiples sources possibles depuis la reconnaissance de la parole jusqu'à la génération en passant par d'autres modules comme l'analyse sémantique, l'interprétation pragmatique ou la gestion du dialogue. Dans cet article, ce problème est envisagé dans le but d'apporter de la généricité et de la robustesse au système ; il est traité au niveau du contrôleur de dialogue. Les différents types d'erreurs sont d'abord identifiés et regroupés en deux catégories qui seules ont un sens vis-à-vis de l'utilisateur : les incompréhensions et les malentendus. Puis, ces deux catégories d'erreur sont traitées de manière spécifique pour que le système puisse générer une réponse convenable et intelligente à l'utilisateur, sans rupture de dialogue. L'expérimentation effectuée en appliquant cette approche au système de dialogue Mélina présente des résultats prometteurs pour traiter les erreurs en dialogue.

Le bon fonctionnement d'Intuition, plate-forme de recherche d'information, repose sur le développement et l'intégration d'un grand nombre de ressources linguistiques. Dans un souci de cohérence et de meilleure gestion, l'unification de ressources contenant des connaissances hétérogènes s'impose. Comme Intuition est disponible dans la plupart des langues européennes, cette unification se heurte au facteur multi-lingue. Pour surmonter les problèmes causés par les différences structurelles entre les langues, une nouvelle architecture linguistique a été conçue et exprimée en UML. Ce méta-modèle est le point de départ pour la nouvelle base de données qui sera le noyau d'un nouvel environnement de travail centré sur son utilisateur, l'expert linguistique. Cet environnement centralisera la gestion de toutes les ressources linguistiques d'Intuition.

Nous avons construit un système capable de reconnaître les modes de composition pour les poèmes arabes, nous décrivons dans cet article les différents modules du système. Le recours à une technique d'apprentissage artificiel pour classer une séquence phonétique de syllabes est justifiable par le fait que nous avons imité le processus d'apprentissage naturel humain suivi par les poètes pendant des siècles. Les réseaux de neurones artificiels de type Perceptron multicouches ont montré un pouvoir très puissant de classification.

La conception de logiciels est un processus technologique complexe, qui nécessite d'être assisté par des outils de traitement automatique des langues. Cet article présente une méthode pour l'annotation de relations discursives de contrôle dans des textes de spécification de besoins informatiques (SBI). La méthode vise à distinguer les actions contrôlées par le système de celles contrôlées par son environnement, ce qui permet d'établir de façon claire les limites et les responsabilités d'un système informatique. Notre méthode fait appel à la sémantique discursive pour analyser les moyens d'expression du contrôle dans un corpus de SBI industrielles ; l'expression du contrôle est identifiable par la présence, dans un certain contexte, de marqueurs linguistiques exprimés par des règles dites d'Exploration Contextuelle. La dernière partie montre le processus d'annotation automatique de la notion de contrôle par le système EXCOM et termine par la présentation d'un début d'évaluation de cette méthodologie.

Nous présentons dans cet article une mémoire de traduction sous-phrastique sensible au domaine de traduction, une première étape vers l'intégration du contexte. Ce système est en mesure de recycler les traductions déjà « vues » par la mémoire, non seulement pour des phrases complètes, mais également pour des sous-séquences contiguës de ces phrases, via un aligneur de mots. Les séquences jugées intéressantes sont proposées au traducteur. Nous expliquons également la création d'un utilisateur artificiel, indispensable pour tester les performances du système en

l'absence d'intervention humaine. Nous le testons lors de la traduction d'un ensemble disparate de corpus. Ces performances sont exprimées par un ensemble de métriques que nous définissons. Enfin, nous démontrons que la détection automatique du contexte de traduction peut s'avérer bénéfique et prometteuse pour améliorer le fonctionnement d'une telle mémoire, en agissant comme un filtre sur le matériel cible suggéré.

Dans ce papier nous proposons d'abord une méthode d'analyse et de désambiguïsation morphologiques de textes arabes non voyellés permettant de lever l'ambiguïté morphologique due à l'absence des marques de voyelles et aussi à l'irrégularité des formes dérivées de certains mots arabes (e.g. formes irrégulières du pluriel des noms et des adjectifs). Ensuite, nous présentons le système MORPH2, un analyseur morphologique de textes arabes non voyellés basé sur la méthode proposée. Ce système est évalué sur un livre scolaire et des articles de journaux. Les résultats obtenus son et très encourageants. En effet, les mesures de rappel et de précision globales sont respectivement de 69,77 % et 68,51 %.

Depuis la conception du Web sémantique une tâche importante se pose au niveau de traitement automatique du langage : rendre accessible le contenu existant duWeb dit classique aux traitements et raisonnements ontologiques. Comme la plupart du contenu est composé de textes, on a besoin de générer des représentations ontologiques de ces informations textuelles. Dans notre article nous proposons une méthode afin d'automatiser cette traduction en utilisant des ontologies et une analyse syntaxico-sémantique profonde.

Dans le domaine du Traitement Automatique du Langage Naturel, pour élaborer un système de représentation thématique des connaissances générales, des méthodes s'appuyant sur des thésaurus sont utilisées depuis une quinzaine d'années. Un thésaurus est constitué d'un ensemble de concepts qui définissent un système générateur d'un espace vectoriel modélisant les

connaissances générales. Ces concepts, souvent organisés en une hiérarchie arborescente, constituent un instrument fondamental, mais totalement figé. Même si les notions évoluent (nous pensons par exemple aux domaines techniques), un thésaurus ne peut quant à lui être modifié que lors d'un processus particulièrement lourd, car nécessitant la collaboration d'experts humains. C'est à ce problème que nous nous attaquons ici. Après avoir détaillé les caractéristiques que doit posséder un système générateur de l'espace vectoriel de modélisation des connaissances, nous définissons les « notions de base ». Celles-ci, dont la construction s'appuie initialement sur les concepts d'un thésaurus, constituent un autre système générateur de cet espace vectoriel. Nous abordons la détermination des acceptions exprimant les notions de base, ce qui nous amène naturellement à nous poser la question de leur nombre. Enfin, nous explicitons comment, s'affranchissant des concepts du thésaurus, ces notions de base évoluent par un processus itératif au fur et à mesure de l'analyse de nouveaux textes.

Dans cet article, nous cherchons à affiner la notion de comparabilité des corpus. Nous étudions en particulier la distinction entre les documents scientifiques et vulgarisés dans le domaine médical. Nous supposons que cette distinction peut apporter des informations importantes, par exemple en recherche d'information. Nous supposons par là même que les documents, étant le reflet de leur contexte de production, fournissent des critères nécessaires à cette distinction. Nous étudions plusieurs critères linguistiques, typographiques, lexicaux et autres pour la caractérisation des documents médicaux scientifiques et vulgarisés. Les résultats présentés sont acquis sur les données en russe et en japonais. Certains des critères étudiés s'avèrent effectivement pertinents. Nous faisons également quelques réflexions et propositions quant à la distinction des catégories scientifique et vulgarisée et aux questionnements théoriques.

La lexicalisation des grammaires réduit le nombre des erreurs d'analyse syntaxique et améliore les résultats des applications. Cependant, cette modification affecte un système d'analyse syntaxique

dans tous ses aspects. Un de nos objectifs de recherche est de mettre au point un modèle réaliste pour la lexicalisation des grammaires. Nous avons réalisé des expériences en ce sens avec une grammaire très simple par son contenu et son formalisme, et un lexique syntaxique très informatif, le lexique-grammaire du français élaboré au LADL. La méthode de lexicalisation est celle des graphes paramétrés. Nos résultats tendent à montrer que la plupart des informations contenues dans le lexique-grammaire peuvent être transférées dans une grammaire et exploitées avec succès dans l'analyse syntaxique de phrases.

Cet article décrit le traitement automatique des pronoms clitiques en espagnol et en grec moderne, deux langues de familles distinctes, dans le cadre de l'analyseur syntaxique FIPS multi-lingue, développé au Laboratoire d'Analyse et de Technologie de Langage (LATL). Nous abordons la distribution des pronoms clitiques, leurs similarités ainsi que leurs particularités par rapport à leur usage général. Ensuite nous présentons la méthode appliquée pour leur traitement, commune aux deux langues. Nous montrons que l'algorithme proposé peut facilement s'étendre à d'autres langues traitées par Fips qui partagent le phénomène de la cliticisation.

Nous présentons dans cet article SIGLé (Système d'Identification de propositions avec Grammaire Légère), un système réalisant la détection des propositions françaises. Ce système détecte les propositions ? à partir de phrases en entrée ségmentées et étiquetées en chunk par un analyseur extérieur ?, analyse leurs relations et leur attribue une étiquette indiquant leur nature syntaxique. Il est caractérisé d'une part par sa grammaire de type CFG proposant un ensemble d'étiquettes adaptées à notre analyse pour les mots dits en « qu- », et d'autre part par l'utilisation du formalisme DCG et du langage PROLOG.

Ce papier présente une nouvelle approche de la segmentation du vietnamien pour la catégorisation de texte. Au lieu d'utiliser des corpus d'entraînement annotés ou des lexiques (qui font défaut pour

le vietnamien) nous utilisons des informations statistiques extraites directement d'un moteur de recherche commercial et des algorithmes génétiques pour trouver les segmentations les plus probables. Les informations extraites incluent la fréquence des documents et l'information mutuelle des n-grams. Nos résultats expérimentaux obtenus sur la segmentation et la catégorisation de résumés de nouvelles montrent que notre approche est très prometteuse. Elle offre des résultats semblables à 80 % avec le jugement humain sur la segmentation et à 90 % en catégorisation. Le temps de traitement est inférieur à une seconde par document quand l'information statistique est maintenue en cache.

Nous présentons, ici, une implémentation d'un système qui n'extrait pas seulement une grammaire lexicalisée (LTAG), mais aussi une grammaire LTAG avec traits (FB-LTAG) à partir d'un corpus arboré. Nous montrons les expérimentations pratiques où nous extrayons les grammaires TAG à partir du Sejong Treebank pour le coréen. Avant tout, les 57 étiquettes syntaxiques et les analyses morphologiques dans le corpus SJTree nous permettent d'extraire les traits syntaxiques automatiquement. De plus, nous modifions le corpus pour l'extraction d'une grammaire lexicalisée et convertissons les grammaires lexicalisées en schémas d'arbre pour résoudre le problème de la couverture lexicale limitée des grammaires lexicalisées extraites.

Cet article présente des expériences récentes menées dans le cadre d'un projet de recherche consacré à l'étude de métaphores conceptuelles. Ces expériences consistent à appréhender visuellement la répartition de trois domaines pouvant être à l'origine de métaphores conceptuelles dans un corpus d'articles boursiers. Les trois domaines étudiés sont la météorologie, la guerre et la santé, un grand nombre d'emplois métaphoriques du lexique de ces trois domaines ayant été observés dans le corpus d'étude. Afin de visualiser la répartition de ces domaines en corpus, nous exploitons la plate-forme ProxiDocs dédiée à la cartographie et à la catégorisation de corpus. Les cartes construites à partir du corpus et des domaines d'étude nous ont ainsi permis de localiser

certaines métaphores conceptuelles dans des articles et des groupes d'articles du corpus. Des articles contenant des emplois non métaphoriques des domaines étudiés ont également été distingués sur les cartes. Des représentations cartographiques du corpus mettant dynamiquement en évidence l'évolution des trois domaines d'étude au fil du temps nous ont permis d'amorcer une étude sur le lien entre la présence de certaines métaphores conceptuelles et des faits d'actualité.

Cet article décrit une approche combinant différents modèles statistiques pour la traduction automatique basée sur les segments. Pour ce faire, différentes ressources sont utilisées, dont deux corpus parallèles aux caractéristiques différentes et un dictionnaire de terminologie bilingue et ce, afin d'améliorer la performance quantitative et qualitative du système de traduction. Nous évaluons notre approche sur la paire de langues français-anglais et montrons comment la combinaison des ressources proposées améliore de façon significative les résultats.

Le lexique grammaire est une méthode systématique d'analyse et de représentation des structures de phrase élémentaire d'une langue naturelle ; son produit : des grandes collections de dictionnaires syntaxiques électroniques ou tables de lexique-grammaire (LGTs). Du travail collaboratif à très long terme est nécessaire pour achever la description d'une langue. Cependant, les outils informatiques de gestion de LGTs actuels ne remplissent pas les besoins suivant : intégration automatique de données multi-source, controle de cohérence de données et de versions, filtrage et tri, formats d'échange, gestion couplée des données et de la documentation, interfaces graphiques (GUIs) dédiées et gestion d'utilisateurs et contrôle d'accès. Dans cet article nous proposons une solution basée sur PostgreSQL et/ou MySQL (systèmes de gestion de bases de données libres), Swing (une librairie pour la programmation de GUIs en Java), JDBC (API pour la connectivité de Java aux bases de données), et StAX (API pour l'analyse et la création des documents en XML).

Nous présentons dans cet article une approche générale pour la modélisation et l'analyse

syntaxique des coordinations elliptiques. Nous montrons que les lexèmes élidés peuvent être remplacés, au cours de l'analyse, par des informations qui proviennent de l'autre membre de la coordination, utilisé comme guide au niveau des dérivations. De plus, nous montrons comment cette approche peut être effectivement mise en oeuvre par une légère extension des Grammaires d'Arbres Adjoints Lexicalisées (LTAG) à travers une opération dite de fusion. Nous décrivons les algorithmes de dérivation nécessaires pour l'analyse de constructions coordonnées pouvant comporter un nombre quelconque d'ellipses.

Nous proposons un cadre théorique qui permet, à partir de matrices construites sur la base des données statistiques d'un corpus, d'extraire par des procédés mathématiques simples des informations sur les mots du vocabulaire de ce corpus, et sur la syntaxe des langues qui l'ont engendré. À partir des mêmes données initiales, on peut construire une matrice de similarité syntagmatique (probabilités de transition d'un mot à un autre), ou une matrice de similarité paradigmatique (probabilité de partager des contextes identiques). Pour ce qui concerne la première de ces deux possibilités, les résultats obtenus sont interprétés dans le cadre d'une modélisation du processus génératif par chaînes de Markov. Nous montrons que les résultats d'une analyse spectrale de la matrice de transition peuvent être interprétés comme des probabilités d'appartenance de mots à des classes. Cette méthode nous permet d'obtenir une classification continue des mots du vocabulaire dans des sous-systèmes génératifs contribuant à la génération de textes composites. Une application pratique est la segmentation de textes hétérogènes en segments homogènes d'un point de vue linguistique, notamment dans le cas de langues proches par le degré de recouvrement de leurs vocabulaires.

Les modèles markoviens de langage sont très dépendants des données d'entraînement sur lesquels ils sont appris. Cette dépendance, qui rend difficile l'interprétation des performances, a surtout un fort impact sur l'adaptation à chaque utilisateur de ces modèles. Cette question a déjà été

largement étudiée par le passé. En nous appuyant sur un domaine d'application spécifique (prédiction de texte pour l'aide à la communication pour personnes handicapées), nous voudrions l'étendre à la problématique de l'influence du registre de langage. En considérant des corpus relevant de cinq genres différents, nous avons étudié la réduction de cette influence par trois modèles adaptatifs différents : (a) un modèle cache classique favorisant les n derniers mots rencontrés, (b) l'intégration au modèle d'un dictionnaire dynamique de l'utilisateur et enfin (c) un modèle de langage interpolé combinant un modèle général et un modèle utilisateur mis à jour dynamiquement au fil des saisies. Cette évaluation porte un système de prédiction de texte basé sur un modèle tri-gramme.

Notre travail s'intègre dans le cadre du projet intitulé « Oréodule » : un système de reconnaissance, de traduction et de synthèse de la langue arabe. L'objectif de cet article est d'essayer d'améliorer le modèle probabiliste sur lequel est basé notre décodeur sémantique de la parole arabe spontanée. Pour atteindre cet objectif, nous avons décidé de tester l'influence de l'utilisation du contexte pertinent, et de l'intégration de différents types de données contextuelles sur la performance du décodeur sémantique employé. Les résultats sont satisfaisants.

On se place ici dans la tendance actuelle en traitement automatique des langues, celle à base de corpus et aussi dans une perspective que l'on peut qualifier d'approche à moindre effort : il s'agit d'examiner les limites des possibilités de traitement à partir de données textuelles brutes, c'est-à-dire non pré-traitées. L'interrogation théorique présente en arrière-plan est la suivante : quelles sont les opérations fondamentales en langue ? L'analogie proportionnelle a été mentionnée par de nombreux grammairiens et linguistes. On se propose de montrer l'efficacité d'une telle opération en la testant sur une tâche dure du traitement automatique des langues : la traduction automatique. On montrera aussi les bonnes conséquences de la formalisation d'une telle opération avec des résultats théoriques en théorie des langages en relation avec leur adéquation à la

description des langues. De cette façon, une opération fondamentale en langue, l'analogie proportionnelle, se verra illustrée tant par ses aspects théoriques que par ses performances en pratique.

Dans cet article, nous nous intéressons à l'interprétation de commandes en langue naturelle pour un agent artificiel. Notre architecture repose sur une modélisation logique de la commande pour l'interprétation sémantique, qui permet de capturer la « structure fonctionnelle » de la phrase, c'est-à-dire les rôles des termes les uns par rapport aux autres. Cet article décrit une méthode d'analyse structurelle de surface qui s'appuie sur l'ontologie de l'agent pour construire cette modélisation logique. Nous définissons tout d'abord un algorithme d'ancrage des termes de la commande dans l'ontologie de l'agent puis nous montrons comment s'en servir pour l'analyse de surface. Enfin, nous expliquons brièvement comment notre modélisation peut être utilisée au moment de l'interprétation sémantique des commandes.

Nous présentons une approche empirique de l'évaluation automatique des réponses d'apprenants au sein d'un système d'Apprentissage des Langues Assisté par Ordinateur (ALAO). Nous proposons la mise en place d'un module d'analyse d'erreurs attestées sur corpus qui s'appuie sur des techniques robustes de Traitement Automatique des Langues (TAL). Cet article montre la réalisation d'un module d'analyse de morphologie flexionnelle, en situation hors-contexte, à partir d'un modèle linguistique existant.

Cet article présente une procédure de repérage et de balisage de l'élément générique de la définition terminographique exploitant les caractéristiques formelles du sous-langage définitoire. La procédure, qui comporte quatre étapes, constitue l'une des sous-tâches d'un analyseur (semi-)automatique de la structure conceptuelle des définitions terminographiques, destiné à faciliter l'annotation d'un corpus en vue de l'étude de régularités dans cette structure. La tâche décrite

consiste à mettre au point un système d'annotation automatique basé sur le repérage d'indices morpho-syntaxiques, sans recourir à d'autres ressources linguistiques informatisées.

Nous décrivons ici comment enrichir automatiquement WordNet en y important des articles encyclopédiques. Ce processus permet de créer des nouvelles entrées, en les rattachant au bon hyperonyme. Par ailleurs, les entrées préexistantes de WordNet peuvent être enrichies de descriptions complémentaires. La répétition de ce processus sur plusieurs encyclopédies permet de constituer un corpus d'articles comparables. On peut ensuite extraire automatiquement des paraphrases à partir des couples d'articles ainsi créés. Grâce à l'application d'une mesure de similarité, utilisant la hiérarchie de verbes de WordNet, les constituants de ces paraphrases peuvent être désambiguïsés.

Dans les ressources dictionnairiques développées à partir du cadre théorique de la Lexicologie Explicative et Combinatoire telles que le DiCo, les relations sémanticolexicales sont modélisées au moyen de fonctions lexicales. Cependant, seulement la majorité d'entre elles (dites standard) répondent véritablement à un encodage formel. Les autres (dites non standard), représentant des relations plus spécifiques à certaines unités lexicales, sont écrites sous la forme d'un encodage hétérogène et très peu formalisé. Par conséquent, certaines relations ne peuvent entrer en ligne de compte dans les traitements automatiques. Nous proposons dans cet article une méthodologie pour la normalisation des fonctions lexicales non standard afin de les rendre exploitables dans des applications telles que l'analyse et la génération de texte. Pour ce faire, nous discutons certains principes théoriques associés à ce formalisme de description et esquissons des propositions pour un traitement global et homogène de l'ensemble des relations décrites dans le DiCo.

Nous nous intéressons aux systèmes multimodaux qui utilisent les modes et modalités suivantes : l'oral (et le langage naturel) en entrée et en sortie, le geste en entrée et le visuel en sortie par

affichage sur écran. L'usager échange avec le système par un geste et/ou un énoncé oral en langue naturelle. Dans cet échange, encodé sur les différentes modalités, se trouvent l'expression du but de l'usager et la désignation des objets (référents) nécessaires à la réalisation de ce but. Le système doit identifier de manière précise et non ambiguë les objets désignés par l'usager. Nous traitons plus spécialement dans cet article les désignations orales, sans geste, des objets dans le contexte visuel. En effet, l'ensemble du contexte multimodal, dont le mode visuel, influe sur la production de l'entrée de l'usager. Afin d'identifier une désignation produite en s'appuyant sur le contexte visuel, nous proposons un algorithme qui utilise des connaissances « classiques » linguistiques, des connaissances sur les objets manipulés, et des connaissances sur les aspects perceptifs (degré de saillance) associés à ces objets.

Cet article propose une évaluation combinée et comparative de 5 ressources (descriptive, paradigmatique et syntagmatiques) pour l'aide à l'accès lexical en situation de "mot sur le bout de la langue", en vue de la création d'un outil utilisant la combinaison de ces ressources. En situation de "mot sur le bout de la langue", l'utilisateur n'accède plus au mot qu'il veut dire ou écrire mais est capable d'en produire d'autres sémantiquement associés. L'évaluation se base sur un corpus de 20 mots "sur le bout de la langue" pour lesquels on dispose de 50 groupes de 5 associations sémantiques effectuées par des utilisateurs. Les résultats montrent que les ressources sont complémentaires et peu redondantes. De plus au moins une association proposée parmi les 5 permettrait de retrouver le mot "sur le bout de la langue" dans 79% des cas, à condition de le sélectionner parmi les 2500 mot potentiels. Enfin, les résultats montrent des disparités entre les utilisateurs, ce qui permettrait de définir des profils d'utilisateur pour une amélioration des performances.

Les systèmes d'Extraction d'Information se contentent, le plus souvent, d'enrichir des bases de données plates avec les informations qu'ils extraient. Nous décrivons dans cet article un travail en

cours sur l'utilisation de données extraites automatiquement pour la construction d'une structure de représentation plus complexe. Cette structure modélise un réseau social composé de relations entre les entités d'un corpus de biographies.

Cet article présente une méthode pour construire, à partir d'une ressource lexicale prédicative existante, une ressource enrichie pouvant servir à une tâche d'extraction. Nous montrons les points forts et les lacunes de deux ressources existantes pour le Français : les Tables du LADL et Volem. Après avoir montré pourquoi nous avons sélectionné Volem, nous listons les données nécessaires à la tâche d'extraction d'information. Nous présentons le processus d'enrichissement de la ressource initiale et une évaluation, à travers une tâche d'extraction d'information concernant des textes de rachats d'entreprise.

Afin de concevoir un agent conversationnel logiciel capable d'assister des utilisateurs novices d'applications informatiques, nous avons été amenés à constituer un corpus spécifique de requêtes d'assistance en français, et à étudier ses caractéristiques. Nous montrons ici que les requêtes d'assistance se distinguent nettement de requêtes issues d'autres corpus disponibles dans des domaines proches. Nous mettons également en évidence le fait que ce corpus n'est pas homogène, mais contient au contraire plusieurs activités conversationnelles distinctes, dont l'assistance elle-même. Ces observations nous permettent de discuter de l'opportunité de considérer l'assistance comme un registre particulier de la langue générale.

Pour des raisons variées, diverses communautés se sont intéressées aux corpus multilingues. Parmi ces corpus, les textes parallèles sont utilisés aussi bien en terminologie, lexicographie ou comme source d'informations pour les systèmes de traduction par l'exemple. L'Union Européenne, qui a entraîné la production de document législatif dans vingtaine de langues, est une des sources de ces textes parallèles. Aussi, avec le Web comme vecteur principal de diffusion de ces textes

parallèles, cet objet d'étude est passé à un nouveau statut : celui de document. Cet article décrit un système d'alignement prenant en compte un grand nombre de langues simultanément (> 2) et les caractéristiques structurelles des documents analysés.

Les campagnes d'évaluation ne tiennent compte que des résultats finaux obtenus par les systèmes de recherche d'informations (RI). Nous nous situons dans une perspective d'évaluation transparente d'un système de questions-réponses, où le traitement d'une question se fait grâce à plusieurs composants séquentiels. Dans cet article, nous nous intéressons à l'étude de l'élément de la question qui porte l'information qui se trouvera dans la phrase réponse à proximité de la réponse elle-même : le focus. Nous définissons ce concept, l'appliquons au système de questions-réponses QALC, et démontrons l'utilité d'évaluations des composants afin d'augmenter la performance globale du système.

Nous faisons l'hypothèse que les bornes délimitées par la méthode statistique TextTiling peuvent servir d'indices qui, cumulées à des indices de nature linguistique, permettront de repérer automatiquement des segments d'informations évolutives. Ce travail est développé dans le cadre d'un projet industriel plus général dont le but est le repérage automatique de zones textuelles contenant de l'information potentiellement évolutive.

Les disfluences (répétitions, amorces, autocorrections, constructions inachevées, etc.) inhérentes à toute production orale spontanée constituent une réelle difficulté en termes d'annotation. En effet, l'annotation de ces phénomènes se révèle difficilement automatisable dans la mesure où leur étude réclame un jugement éminemment interprétatif. Dans cet article, nous présentons une méthodologie applicable à l'annotation des disfluences (ou « phénomènes de production ») que l'on rencontre fréquemment dans les corpus oraux. Le fait de constituer un tel corpus de données annotées, permet non seulement de représenter certains aspects pertinents de l'oral (de manière à servir de

base aux observations et aux comparaisons avec d'autres données) mais aussi d'améliorer in fine le traitement automatique de l'oral (notamment l'analyse syntaxique automatique).

La génération du langage naturel pour le dialogue oral homme-machine pose des contraintes spécifiques, telles que la spontanéité et le caractère fragmenté des énoncés, les types des locuteurs ou les contraintes de temps de réponse de la part du système. Dans ce contexte, le problème d'une architecture rigoureusement spécifiée se pose, autant au niveau des étapes de traitement et des modules impliqués, qu'au niveau des interfaces entre ces modules. Afin de permettre une liberté quasi-totale à l'égard des démarches théoriques, une telle architecture doit être à la fois modulaire (c'est-à-dire, permettre l'indépendance des niveaux de traitement les uns des autres) et portable (c'est-à-dire, permettre l'interopérabilité avec des modules conçus selon des architectures standard en génération du langage naturel, telles que le modèle RAGS - « Reference Architecture for Generation Systems »). Ainsi, dans cet article on présente de manière concise l'architecture proposée, la comparant ensuite au modèle RAGS, pour argumenter les choix opérés en conception. Dans un second temps, la portabilité de l'architecture sera décrite à travers un exemple étendu, dont la généralité réside dans l'obtention d'un ensemble de règles permettant de plonger automatiquement les représentations des informations de notre architecture vers le format du modèle RAGS et inversement. Finalement, un ensemble de conclusions et perspectives clôturera l'article.

Nous présentons une résolution anaphorique intégrée à une analyse automatique de discours. Cette étude traite des anaphores pronominales et des anaphores zéro. Notre analyse est basée sur trois approches : une analyse basée sur les contraintes, une analyse fonctionnelle et une analyse dynamique. Pour évaluer la faisabilité et la fiabilité de notre approche, nous l'avons expérimentée sur un corpus de 97 histoires produites à l'oral par des enfants. Nous présentons le résultat de cette évaluation.

Antidote RX est la sixième édition d'Antidote, un logiciel d'aide à la rédaction développé et commercialisé par la société Druide informatique. Antidote RX comporte un correcteur grammatical avancé, dix dictionnaires de consultation et dix guides linguistiques. Il fonctionne sous les systèmes d'exploitation Windows, Mac OS X et Linux.

Cordial est un correcteur efficace et discret enrichi d'un grand nombre de fonctions d'aide à la rédaction et d'analyse de documents. Très riche avec ces multiples dictionnaires et souvent pertinent dans ses propositions, Cordial est un compagnon précieux qui vous permet d'assurer la qualité de vos écrits. La version 2007 de Cordial s'intègre dans un vaste éventail de logiciels comme les traitements de texte (Word, Open Office, Word Perfect...), clients de messagerie (Outlook, Notes, Thunderbird, webmails...) ou navigateurs (Explorer, Mozilla).

Nous offrirons une démonstration de la dernière version de TransCheck, un vérificateur automatique de traductions que le RALI est en train de développer. TransCheck prend en entrée deux textes, un texte source dans une langue et sa traduction dans une autre, les aligne au niveau de la phrase et ensuite vérifie les régions alignées pour s'assurer de la présence de certains équivalents obligatoires (p. ex. la terminologie normalisée) et de l'absence de certaines interdictions de traduction (p. ex. des interférences de la langue source). Ainsi, TransCheck se veut un nouveau type d'outil d'aide à la traduction qui pourra à réduire le fardeau de la révision et diminuer le coût du contrôle de la qualité.

Créé en 2005 à l'initiative du Centre National de la Recherche Scientifique, le CNRTL propose une plate-forme unifiée pour l'accès aux ressources et documents électroniques destinés à l'étude et l'analyse de la langue française. Les services du CNRTL comprennent le recensement, la documentation (méta-données), la normalisation, l'archivage, l'enrichissement et la diffusion des

ressources. La pérennité du service et des données est garantie par le soutien institutionnel du CNRS, l'adossement à un laboratoire de recherche en linguistique et informatique du CNRS et de Nancy Université (ATILF ? Analyse et Traitement Informatique de la Langue Française), ainsi que l'intégration dans le réseau européen CLARIN (common language resources and technology infrastructure european).

Dans cet article, nous traitons de la segmentation automatique des textes en épisodes thématiques non superposés et ayant une structure linéaire. Notre étude porte sur l'utilisation des traits lexicaux, acoustiques et syntaxiques et sur l'influence de ces traits sur la performance d'un système automatique de segmentation thématique. Nous appliquons notre approche, basée sur des machines à vecteurs support, à des transcriptions des dialogues mult-ilocuteurs.

Dans cet article, nous présentons une approche de réseaux de neurones inspirée de la physique statistique de systèmes magnétiques pour étudier des problèmes fondamentaux du Traitement Automatique de la Langue Naturelle. L'algorithme modélise un document comme un système de neurones où l'on déduit l'énergie textuelle. Nous avons appliqué cette approche aux problèmes de résumé automatique et de détection de frontières thématiques. Les résultats sont très encourageants.

Dans cet article, nous présentons une méthode permettant d'extraire à partir de textes des relations sémantiques dans le domaine médical en utilisant des patrons linguistiques. La première partie de cette méthode consiste à identifier les entités entre lesquelles les relations visées interviennent, en l'occurrence les maladies, les examens, les médicaments et les symptômes. La présence d'une des relations sémantiques visées dans les phrases contenant un couple de ces entités est ensuite validée par l'application de patrons linguistiques préalablement appris de manière automatique à partir d'un corpus annoté. Nous rendons compte de l'évaluation de cette méthode sur un corpus en

Français pour quatre relations.

On oppose souvent en TAL les systèmes à base de connaissances linguistiques et ceux qui reposent sur des indices de surface. Chaque approche a ses limites et ses avantages. Nous proposons dans cet article une nouvelle approche qui repose sur les réseaux bayésiens et qui permet de combiner au sein d'une même représentation ces deux types d'informations hétérogènes et complémentaires. Nous justifions l'intérêt de notre approche en comparant les performances du réseau bayésien à celles des systèmes de l'état de l'art, sur un problème difficile du TAL, celui de la résolution d'anaphore.

Dans cet article, nous présentons différentes contraintes mécaniques et linguistiques applicables à des règles d'analyse des mots inconnus afin d'améliorer la performance d'un analyseur morphologique de l'italien. Pour mesurer l'impact de ces contraintes, nous présentons les résultats d'une évaluation de chaque contrainte qui prend en compte les gains et les pertes qu'elle engendre. Nous discutons ainsi de la nécessaire évaluation de chaque réglage apporté aux règles afin d'en déterminer la pertinence.

Les structures de traits typées sont une façon abstraite et agréable de représenter une information partielle. Dans cet article, nous montrons comment la combinaison de deux techniques relativement classiques permet de définir une variante de morphologie à deux niveaux intégrant harmonieusement des structures de traits et se compilant en une machine finie. La première de ces techniques est la compilation de structure de traits en expressions régulières, la seconde est la morphologie à partition. Nous illustrons au moyen de deux exemples l'expressivité d'un formalisme qui rapproche les grammaires à deux niveaux des grammaires d'unification.

La plupart des vocabulaires spécialisés comprennent une part importante de lexèmes

morphologiquement complexes, construits à partir de racines grecques et latines, qu'on appelle « composés savants ». Une analyse morphosémantique permet de décomposer et de donner des définitions à ces lexèmes, et semble pouvoir être appliquée de façon similaire aux composés de plusieurs langues. Cet article présente l'adaptation d'un analyseur morphosémantique, initialement dédié au français (DériF), à l'analyse de composés savants médicaux anglais, illustrant ainsi la similarité de structure de ces composés dans des langues européennes proches. Nous exposons les principes de cette transposition et ses performances. L'analyseur a été testé sur un ensemble de 1299 lexèmes extraits de la terminologie médicale WHO-ART : 859 ont pu être décomposés et définis, dont 675 avec succès. Outre une simple transposition d'une langue à l'autre, la méthode montre la potentialité d'un système multi-lingue.

Les congénères sont des mots qui ont au moins un sens en commun entre deux langues en plus d?avoir une orthographie semblable. La reconnaissance de ce type de mots permet aux apprenants de langue seconde ou étrangère d?enrichir plus rapidement leur vocabulaire et d?améliorer leur compréhension écrite. Toutefois, les faux amis sont des paires de mots qui à l?écrit ont des similarités, mais ils ont des significations différentes. Pour leur part, les congénères partiels sont des mots qui ont la même signification dans certains contextes dans chacune des deux langues. Cet article présente une méthode pour la classification automatique des paires des mots classées en congénères ou faux amis, en utilisant des mesures de similarité orthographiques et des méthodes d?apprentissage automatique. Ainsi, nous construisons des listes complètes des congénères et des faux amis entre les deux langues. Nous désambiguisons les congénères partiels dans des contextes spécifiques. Nos méthodes sont évaluées pour le français et l?anglais, mais elles seraient applicables à d?autres paires des langues. Nous avons construit un outil qui prend ces listes et marque dans un texte français les mots qui ont des congénères ou des faux amis en anglais, dans le but d?aider les apprenants en français langue seconde ou étrangère à améliorer leur compréhension écrite et à développer une meilleure rétention.

La présence de mots inconnus dans les applications langagières représente un défi de taille bien connu auquel n'échappe pas la traduction automatique. Les systèmes professionnels de traduction offrent à cet effet à leurs utilisateurs la possibilité d'enrichir un lexique de base avec de nouvelles entrées. Récemment, Stroppa & Yvon (2005) démontraient l'intérêt du raisonnement par analogie pour l'analyse morphologique d'une langue. Dans cette étude, nous montrons que le raisonnement par analogie offre également une réponse adaptée au problème de la traduction d'entrées lexicales inconnues.

Dans le domaine biomédical, le caractère multi-lingue de l'accès à l'information est un problème d'importance. Dans cet article nous présentons une technique originale permettant de traduire des termes simples du domaine biomédical de et vers de nombreuses langues. Cette technique entièrement automatique repose sur l'apprentissage de règles de réécriture à partir d'exemples et l'utilisation de modèles de langues. Les évaluations présentées sont menées sur différentes paires de langues (français-anglais, espagnol-portugais, tchèque-anglais, russe-anglais...). Elles montrent que cette approche est très efficace et offre des performances variables selon les langues mais très bonnes dans l'ensemble et nettement supérieures à celles disponibles dans l'état de l'art. Les taux de précision de traductions s'étagent ainsi de 57.5% pour la paire russe-anglais jusqu'à 85% pour la paire espagnol-portugais et la paire français-anglais.

Nous présentons le logiciel TiLT pour la correction des SMS et évaluons ses performances sur le corpus de SMS du DELIC. L'évaluation utilise la distance de Jaccard et la mesure BLEU. La présentation des résultats est suivie d'une analyse qualitative du système et de ses limites.

Un "méta-EDL" (méta-Environnement de Développement Linguiciel) pour la TAO permet de piloter à distance un ou plusieurs EDL pour construire des systèmes de TAO hétérogènes. Partant de CASH,

un méta-EDL dédié à Ariane-G5, et de WICALE 1.0, un premier méta-EDL générique mais aux fonctionnalités minimales, nous dégageons les problèmes liés à l'ajout de fonctionnalités riches comme l'édition et la navigation en local, et donnons une solution implémentée dans WICALE 2.0. Nous y intégrons maintenant une base lexicale pour les systèmes à « pivot lexical », comme UNL/U++. Un but à plus long terme est de passer d'un tel méta-EDL générique multifonctionnel à un EDL « universel », ce qui suppose la réingénierie des compilateurs et des moteurs des langages spécialisés pour la programmation linguistique (LSPL) supportés par les divers EDL.

Dans cet article, nous évaluons les performances de fonctionnalités d'aide à la navigation dans un contexte de recherche dans un corpus audio. Nous montrons que les particularités de la transcription et, en particulier les erreurs, conduisent à une dégradation parfois importante des performances des outils d'analyse. Si la navigation par concepts reste dans des niveaux d'erreur acceptables, la reconnaissance des entités nommées, utilisée pour l'aide à la lecture, voit ses performances fortement baisser. Notre remise en doute de la portabilité de ces fonctions à un corpus oral est néanmoins atténuée par la nature même du corpus qui incite à considérer que toute méthodes permettant de réduire le temps d'accès à l'information est pertinente, même si les outils utilisés sont imparfaits.

Dans cet article, nous présentons une grammaire du français qui fait l'objet d'un modèle basé sur des descriptions linguistiques de corpus (provenant notamment des travaux de l'Approche Pronominale) et représentée selon le formalisme des Grammaires de Propriétés. Elle constitue une proposition nouvelle parmi les grammaires formelles du français, participant à la mise en convergence de la variété des travaux de description linguistique, et de la diversité des possibilités de représentation formelle. Cette grammaire est mise à disposition publique sur le Centre de Ressources pour la Description de l'Oral en tant que ressource pour la représentation et l'analyse.

L'article présente les principes généraux sous-jacent aux grammaires catégorielles de dépendances : une classe de grammaires de types récemment proposée pour une description compositionnelle et uniforme des dépendances continues et discontinues. Ces grammaires très expressives et analysées en temps polynomial, adoptent naturellement l'architecture multimodale et expriment les dépendances croisées illimitées.

Dans cet article, nous présentons une architecture logicielle libre et ouverte pour le développement de grammaires d'arbres adjoints à portée sémantique. Cette architecture utilise un compilateur de méta-grammaires afin de faciliter l'extension et la maintenance de la grammaire, et intègre un module de construction sémantique permettant de vérifier la couverture aussi bien syntaxique que sémantique de la grammaire. Ce module utilise un analyseur syntaxique tabulaire généré automatiquement à partir de la grammaire par le système DyALog. Nous présentons également les résultats de l'évaluation d'une grammaire du français développée au moyen de cette architecture.

La désambiguïsation lexicale présente un intérêt considérable pour un nombre important d'applications, en traitement automatique des langues comme en recherche d'information. Nous proposons un modèle d'un genre nouveau, fondé sur la théorie de la construction dynamique du sens (Victorri et Fuchs, 1996). Ce modèle donne une place centrale à la polysémie et propose une représentation géométrique du sens. Nous présentons ici une application de ce modèle à la désambiguïsation automatique des adjectifs. La méthode utilisée s'appuie sur une pré-désambiguïsation du nom régissant l'adjectif, par le biais de classes de sélection distributionnelle. Elle permet aussi de prendre en compte les positions relatives du nom et de l'adjectif (postpostion ou antéposition) dans le calcul du sens.

La relation voir/employé pour d'un thesaurus est souvent plus complexe que la (para-)synonymie recommandée par l'ISO-27-88, standard décrivant le contenu de ces vocabulaires contrôlés. Le fait

qu'un non descripteur puisse renvoyer à plusieurs descripteurs (seuls les descripteurs sont pertinents dans le cadre de l'indexation contrôlée) fait que cette relation est complexe à utiliser dans un contexte d'annotation automatique : elle génère des cas d'ambiguité. Dans ce papier, nous présentons CARROT, un algorithme que nous avons mis au point pour classer les résultats de notre chaîne de traitements pour l'Extraction d'Information, et son utilisation dans le cadre de la sélection du descripteur pertinent lorsque plusieurs choix sont possibles. Cette sélection s'adresse à des documentalistes, dans le but de simplifier et d'accélérer leur travail, et se base sur la structure de leur thesaurus. Nous arrivons à un succès de 95 % dans nos suggestions ; nous discutons ces résultats et présentons des perspectives à cette expérimentation.

Les besoins de désambiguïsation varient dans les différentes applications du Traitement Automatique des Langues (TAL). Dans cet article, nous proposons une méthode de désambiguïsation lexicale opératoire dans un contexte bilingue et, par conséquent, adéquate pour la désambiguïsation au sein d'applications relatives à la traduction. Il s'agit d'une méthode contextuelle, qui combine des informations de co-occurrence avec des informations traductionnelles venant d'un bi-texte. L'objectif est l'établissement de correspondances de traduction au niveau sémantique entre les mots de deux langues. Cette méthode étend les conséquences de l'hypothèse contextuelle du sens dans un contexte bilingue, tout en admettant l'existence d'une relation de similarité sémantique entre les mots de deux langues en relation de traduction. La modélisation de ces correspondances de granularité fine permet la désambiguïsation lexicale de nouvelles occurrences des mots polysémiques de la langue source ainsi que la prédiction de la traduction la plus adéquate pour ces occurrences.

PrepLex est un lexique des prépositions du français. Il contient les informations utiles à des systèmes d'analyse syntaxique. Il a été construit en comparant puis fusionnant différentes sources d'informations lexicales disponibles. Ce lexique met également en évidence les prépositions ou

classes de prépositions qui apparaissent dans la définition des cadres de sous-catégorisation des ressources lexicales qui décrivent la valence des verbes.

Cet article compare le Lexique-Grammaire des verbes pleins et DICOVALENCE, deux ressources lexicales syntaxiques pour le français développées par des linguistes depuis de nombreuses années. Nous étudions en particulier les divergences et les empiètements des modèles lexicaux sous-jacents. Puis nous présentons le Lefff , lexique syntaxique à grande échelle pour le TAL, et son propre modèle lexical. Nous montrons que ce modèle est à même d'intégrer les informations lexicales présentes dans le Lexique-Grammaire et dans DICOVALENCE. Nous présentons les résultats des premiers travaux effectués en ce sens, avec pour objectif à terme la constitution d'un lexique syntaxique de référence pour le TAL.

Nous présentons dans cet article le prototype d'un système d'étiquetage syntactico-sémantique des mots qui utilise comme principales ressources linguistiques différents dictionnaires du laboratoire Lexiques, Dictionn-aires, Informatique (LDI). Dans un premier temps, nous mentionnons des travaux sur le même sujet. Dans un deuxième temps, nous faisons la présentation générale du système. Dans un troisième temps, nous exposons les principales caractéristiques des dictionnaires syntactico-sémantiques utilisés. Dans un quatrième temps, nous détaillons un exemple de traitement.

Cet article s'intéresse au problème de la détection et de la correction des erreurs cachées sémantiques dans les textes arabes. Ce sont des erreurs orthographiques produisant des mots lexicalement valides mais invalides sémantiquement. Nous commençons par décrire le type d'erreur sémantique auquel nous nous intéressons. Nous exposons par la suite l'approche adoptée qui se base sur la combinaison de plusieurs méthodes, tout en décrivant chacune de ces méthodes. Puis, nous évoquons le contexte du travail qui nous a mené au choix de l'architecture multi-agent pour

l'implémentation de notre système. Nous présentons et commentons vers la fin les résultats de l'évaluation dudit système.

Cet article décrit deux approches, l'une numérique, l'autre symbolique, traitant le problème de la résolution de la référence dans un cadre de dialogue homme-machine. L'analyse des résultats obtenus sur le corpus MEDIA montre la complémentarité des deux systèmes développés : robustesse aux erreurs et hypothèses multiples pour l'approche numérique ; modélisation de phénomènes complexes et interprétation complète pour l'approche symbolique.

Dans le paradigme FrameNet, cet article aborde le problème de l'annotation précise et automatique de rôles sémantiques dans une langue sans lexique FrameNet existant. Nous évaluons la méthode proposée par Padó et Lapata (2005, 2006), fondée sur la projection de rôles et appliquée initialement à la paire anglais-allemand. Nous testons sa généralisabilité du point de vue (a) des langues, en l'appliquant à la paire (anglais-français) et (b) de la qualité de la source, en utilisant une annotation automatique du côté anglais. Les expériences montrent des résultats à la hauteur de ceux obtenus pour l'allemand, nous permettant de conclure que cette approche présente un grand potentiel pour réduire la quantité de travail nécessaire à la création de telles ressources dans de nombreuses langues.

Antidote RX, un logiciel d'aide à la rédaction grand public, comporte un nouveau dictionnaire de 800 000 co-occurrences, élaboré essentiellement automatiquement. Nous l'avons créé par l'analyse syntaxique détaillée d'un vaste corpus et par la sélection automatique des co-occurrences les plus pertinentes à l'aide d'un test statistique, le rapport de vraisemblance. Chaque co-occurrence est illustrée par des exemples de phrases également tirés du corpus automatiquement. Les co-occurrences et les exemples extraits ont été révisés par des linguistes. Nous examinons les choix d'interface que nous avons faits pour présenter ces données complexes à un public non

spécialisé. Enfin, nous montrons comment nous avons intégré les co-occurrences au correcteur d'Antidote pour améliorer ses performances.

Fréquemment utilisés dans le Traitement Automatique des Langues Naturelles, les réseaux lexicaux font aujourd'hui l'objet de nombreuses recherches. La plupart d'entre eux, et en particulier le plus célèbre Word-Net, souffrent du manque d'informations syntagmatiques mais aussi d'informations thématiques (« problème du tennis »). Cet article présente les vecteurs conceptuels qui permettent de représenter les idées contenues dans un segment textuel quelconque et permettent d'obtenir une vision continue des thématiques utilisées grâce aux distances calculables entre eux. Nous montrons leurs caractéristiques et en quoi ils sont complémentaires des réseaux lexico-sémantiques. Nous illustrons ce propos par l'enrichissement des données de Word-Net par des vecteurs conceptuels construits par émergence.

Ce travail présente une application d'alignement monolingue qui répond à une problématique posée par la critique génétique textuelle, une école d'études littéraires qui s'intéresse à la genèse textuelle en comparant les différentes versions d'une oeuvre. Ceci nécessite l'identification des déplacements, cependant, le problème devient ainsi NP-complet. Notre algorithme heuristique est basé sur la reconnaissance des homologies entre séquences de caractères. Nous présentons une validation expérimentale et montrons que notre logiciel obtient de bons résultats ; il permet notamment l'alignement de livres entiers.

Le succès de l'analyse syntaxique d'une phrase dépend de la qualité de la grammaire sous-jacente mais aussi de celle du lexique utilisé. Une première étape dans l'amélioration des lexiques consiste à identifier les entrées lexicales potentiellement erronées, par exemple en utilisant des techniques de fouilles d'erreurs sur corpus (Sagot & Villemonte de La Clergerie, 2006). Nous explorons ici l'étape suivante : la suggestion de corrections pour les entrées identifiées. Cet objectif est atteint au

travers de réanalyses des phrases rejetées à l'étape précédente, après modification des informations portées par les entrées suspectées. Un calcul statistique sur les nouveaux résultats permet ensuite de mettre en valeur les corrections les plus pertinentes.

En s'appuyant sur la notion d'arbre de dérivation des Grammaires d'Arbres Adjoints (TAG), cet article propose deux objectifs : d'une part rendre l'interface entre syntaxe et sémantique indépendante du langage de représentation sémantique utilisé, et d'autre part offrir un noyau qui permette le traitement sémantique des ambiguïtés de portée de quantificateurs sans utiliser de langage de représentation sous-spécifiée.

SYNLEX est un lexique syntaxique extrait semi-automatiquement des tables du LADL. Comme les autres lexiques syntaxiques du français disponibles et utilisables pour le TAL (LEFFF, DICOVALENCE), il est incomplet et n'a pas fait l'objet d'une évaluation permettant de déterminer son rappel et sa précision par rapport à un lexique de référence. Nous présentons une approche qui permet de combler au moins partiellement ces lacunes. L'approche s'appuie sur les méthodes mises au point en acquisition automatique de lexique. Un lexique syntaxique distinct de SYNLEX est acquis à partir d'un corpus de 82 millions de mots puis utilisé pour valider et compléter SYNLEX. Le rappel et la précision de cette version améliorée de SYNLEX sont ensuite calculés par rapport à un lexique de référence extrait de DICOVALENCE.

Comment produire de façon massive des textes annotés dans des conditions d'efficacité, de reproductibilité et de coût optimales ? Plutôt que de corriger les sorties d'analyse automatique moyennant des outils d'éditions éventuellement dédiés, ainsi qu'il estcommunément préconisé, nous proposons de recourir à des outils d'analyse interactive où la correction manuelle est au fur et à mesure prise en compte par l'analyse automatique. Posant le problème de l'évaluation de ces outils interactifs et du rendement de leur ergonomie linguistique, et proposant pour cela une métrique

fondée sur le calcul du coût qu'exigent ces corrections exprimé en nombre de manipulations (frappe au clavier, clic de souris, etc.), nous montrons, au travers d'un protocole expérimental simple orienté vers la voyellation, l'étiquetage et la lemmatisation de l'arabe, que paradoxalement, les meilleures performances interactives d'un système ne sont pas toujours corrélées à ses meilleures performances automatiques. Autrement dit, que le comportement linguistique automatique le plus performant n'est pas toujours celui qui assure, dès lors qu'il y a contributions manuelles, le meilleur rendement interactif.

Cet article décrit un système pour définir et évaluer les stades de développement en français langue étrangère. L'évaluation de tels stades correspond à l'identification de la fréquence de certains phénomènes lexicaux et grammaticaux dans la production des apprenants et comment ces fréquences changent en fonction du temps. Les problèmes à résoudre dans cette démarche sont triples : identifier les attributs les plus révélateurs, décider des points de séparation entre les stades et évaluer le degré d'efficacité des attributs et de la classification dans son ensemble. Le système traite ces trois problèmes. Il se compose d'un analyseur morpho-syntaxique, appelé Direkt Profil, auquel nous avons relié un module d'apprentissage automatique. Dans cet article, nous décrivons les idées qui ont conduit au développement du système et son intérêt. Nous présentons ensuite le corpus que nous avons utilisé pour développer notre analyseur morpho-syntaxique. Enfin, nous présentons les résultats sensiblement améliorés des classificateurs comparé aux travaux précédents (Granfeldt et al., 2006). Nous présentons également une méthode de sélection de paramètres afin d'identifier les attributs grammaticaux les plus appropriés.

Cet article présente un système d'acquisition de familles morphologiques qui procède par apprentissage non supervisé à partir de listes de mots extraites de corpus de textes. L'approche consiste à former des familles par groupements successifs, similairement aux méthodes de classification ascendante hiérarchique. Les critères de regroupement reposent sur la similarité

graphique des mots ainsi que sur des listes de préfixes et de paires de suffixes acquises automatiquement à partir des corpus traités. Les résultats obtenus pour des corpus de textes de spécialité en français et en anglais sont évalués à l'aide de la base CELEX et de listes de référence construites manuellement. L'évaluation démontre les bonnes performances du système, indépendamment de la langue, et ce malgré la technicité et la complexité morphologique du vocabulaire traité.

L'apprentissage non supervisé permet la découverte de catégories initialement inconnues. Les techniques actuelles permettent d'explorer des séquences de phénomènes alors qu'on a tendance à se focaliser sur l'analyse de phénomènes isolés ou sur la relation entre deux phénomènes. Elles offrent ainsi de précieux outils pour l'analyse de données organisées en séquences, et en particulier, pour la découverte de structures textuelles. Nous présentons ici les résultats d'une première tentative de les utiliser pour inspecter les suites de verbes provenant de phrases de récits d'accident de la route. Les verbes étaient encodés comme paires (cat, temps), où cat représente la catégorie aspectuelle d'un verbe, et temps son temps grammatical. L'analyse, basée sur une approche originale, a fourni une classification des enchaînements de deux verbes successifs en quatre groupes permettant de segmenter les textes. Nous donnons ici une interprétation de ces groupes à partir de statistiques sur des annotations sémantiques indépendantes.

Nous proposons D-STAG, un formalisme pour le discours qui utilise les TAG synchrones. Les analyses sémantiques produites par D-STAG sont des structures de discours hiérarchiques annotées de relations de discours coordonnantes ou subordonnantes. Elles sont compatibles avec les structures de discours produites tant en RST qu'en SDRT. Les relations de discours coordonnantes et subordonnantes sont modélisées respectivement par les opérations de substitution et d'adjonction introduites en TAG.

Bien que de nombreux efforts aient été déployés pour extraire des collocations à partir de corpus de textes, seule une minorité de travaux se préoccupent aussi de rendre le résultat de l'extraction prêt à être utilisé dans les applications TAL qui pourraient en bénéficier, telles que la traduction automatique. Cet article décrit une méthode précise d'identification de la traduction des collocations dans un corpus parallèle, qui présente les avantages suivants : elle peut traiter des collocation flexibles (et pas seulement figées) ; elle a besoin de ressources limitées et d'un pouvoir de calcul raisonnable (pas d'alignement complet, pas d'entraînement) ; elle peut être appliquée à plusieurs paires des langues et fonctionne même en l'absence de dictionnaires bilingues. La méthode est basée sur l'information syntaxique provenant du parseur multi-lingue Fips. L'évaluation effectuée sur 4000 collocations de type verbe-objet correspondant à plusieurs paires de langues a montré une précision moyenne de 89.8% et une couverture satisfaisante (70.9%). Ces résultats sont supérieurs à ceux enregistrés dans l'évaluation d'autres méthodes de traduction de collocations.

L'alignement de phrases à partir de textes bilingues consiste à reconnaître les phrases qui sont traductions les unes des autres. Cet article présente une nouvelle approche pour aligner les phrases d'un corpus parallèle. Cette approche est basée sur la recherche cross-lingue d'information et consiste à construire une base de données des phrases du texte cible et considérer chaque phrase du texte source comme une requête à cette base. La recherche cross-lingue utilise un analyseur linguistique et un moteur de recherche. L'analyseur linguistique traite aussi bien les documents à indexer que les requêtes et produit un ensemble de lemmes normalisés, un ensemble d'entités nommées et un ensemble de mots composés avec leurs étiquettes morpho-syntaxiques. Le moteur de recherche construit les fichiers inversés des documents en se basant sur leur analyse linguistique et retrouve les documents pertinents à partir de leur indexes. L'aligneur de phrases a été évalué sur un corpus parallèle Arabe-Français et les résultats obtenus montrent que 97% des phrases ont été correctement alignées.

Nous exposons dans cet article une expérience de sélection automatique des indices du contexte pour la désambiguïsation lexicale automatique. Notre point de vue est qu'il est plus judicieux de privilégier la pertinence des indices du contexte plutôt que la sophistication des algorithmes de désambiguïsation utilisés. La sélection automatique des indices par le biais d'un algorithme génétique améliore significativement les résultats obtenus dans nos expériences précédentes tout en confortant des observations que nous avions faites sur la nature et la répartition des indices les plus pertinents.

Nous proposons d'exposer ici une méthodologie d'analyse et de représentation d'une des composantes de la structuration des textes, celle liée à la notion de prise en charge énonciative. Nous mettons l'accent sur la structure hiérarchisée des segments textuels qui en résulte ; nous la représentons d'une part sous forme d'arbre et d'autre part sous forme de graphe. Ce dernier permet d'appréhender la dynamique énonciative et modale de textes comme un cheminement qui s'opère entre différents niveaux de discours dans un texte au fur et à mesure de sa lecture syntagmatique.

Depuis l'analyseur développé par Harris à la fin des années 50, les unités polylexicales ont peu à peu été intégrées aux analyseurs syntaxiques. Cependant, pour la plupart, elles sont encore restreintes aux mots composés qui sont plus stables et moins nombreux. Toutefois, la langue est remplie d'expressions semi-figées qui forment également des unités sémantiques : les expressions adverbiales et les collocations. De même que pour les mots composés traditionnels, l'identification de ces structures limite la complexité combinatoire induite par l'ambiguïté lexicale. Dans cet article, nous détaillons une expérience qui intègre ces notions dans un processus de segmentation en super-chunks, préalable à l'analyse syntaxique. Nous montrons que notre chunker, développé pour le français, atteint une précision et un rappel de 92,9 % et 98,7 %, respectivement. Par ailleurs, les unités polylexicales réalisent 36,6 % des attachements internes aux constituants nominaux et prépositionnels.

Nous étudions le rôle des entités nommées et marques discursives de rétroaction pour la tâche de classification et prédiction de la satisfaction usager à partir de dialogues. Les expériences menées sur 1027 dialogues Personne-Machine dans le domaine des agences de voyage montrent que les entités nommées et les marques discursives n'améliorent pas de manière significative le taux de classification des dialogues. Par contre, elles permettent une meilleure prédiction de la satisfaction usager à partir des premiers tours de parole usager.

Dans tout dialogue, les phrases elliptiques sont très nombreuses. Dans cet article, nous évaluons leur impact sur la reconnaissance et la traduction dans le système de traduction automatique de la parole MedSLT. La résolution des ellipses y est effectuée par une méthode robuste et portable, empruntée aux systèmes de dialogue homme-machine. Cette dernière exploite une représentation sémantique plate et combine des techniques linguistiques (pour construire la représentation) et basées sur les exemples (pour apprendre sur la base d'un corpus ce qu'est une ellipse bien formée dans un sous-domaine donné et comment la résoudre).

Cette étude présente la problématique de l'analyse automatique de sondages téléphoniques d'opinion. Cette analyse se fait en deux étapes : tout d'abord extraire des messages oraux les expressions subjectives relatives aux opinions de utilisateurs sur une dimension particulière (efficacité, accueil, etc.) ; puis sélectionner les messages fiables, selon un ensemble de mesures de confiance, et estimer la distribution des diverses opinions sur le corpus de test. Le but est d'estimer une distribution aussi proche que possible de la distribution de référence. Cette étude est menée sur un corpus de messages provenant de vrais utilisateurs fournis par France Télécom R&D.

En génération, un réalisateur de surface a pour fonction de produire, à partir d'une représentation conceptuelle donnée, une phrase grammaticale. Les réalisateur existants soit utilisent une

grammaire réversible et des méthodes statistiques pour déterminer parmi l'ensemble des sorties produites la plus plausible ; soit utilisent des grammaires spécialisées pour la génération et des méthodes symboliques pour déterminer la paraphrase la plus appropriée à un contexte de génération donné. Dans cet article, nous présentons GENI, un réalisateur de surface basé sur une grammaire d'arbres adjoints pour le français qui réconcilie les deux approches en combinant une grammaire réversible avec une sélection symbolique des paraphrases.

Cet article revient sur le type particulier des questions définitoires étudiées dans le cadre des campagnes d'évaluation des systèmes de Questions/Réponses. Nous présentons l'approche développée suite à notre participation à la campagne EQueR et son évaluation lors de QA@CLEF 2006. La réponse proposée est la plus représentative des expressions présentes en apposition avec l'objet à définir, sa sélection est faite depuis des indices dérivés de ces appositions. Environ 80% de bonnes réponses sont trouvées sur les questions définitoires des volets francophones de CLEF. Les cas d'erreurs rencontrés sont analysés et discutés en détail.

L'objectif principal de notre travail consiste à étudier la notion de comparabilité des corpus, et nous abordons cette question dans un contexte monolingue en cherchant à distinguer les documents scientifiques et vulgarisés. Nous travaillons séparément sur des corpus composés de documents du domaine médical dans trois langues à forte distance linguistique (le français, le japonais et le russe). Dans notre approche, les documents sont caractérisés dans chaque langue selon leur thématique et une typologie discursive qui se situe à trois niveaux de l'analyse des documents : structurel, modal et lexical. Le typage des documents est implémenté avec deux algorithmes d'apprentissage (SVMlight et C4.5). L'évaluation des résultats montre que la typologie discursive proposée est portable d'une langue à l'autre car elle permet en effet de distinguer les deux discours. Nous constatons néanmoins des performances très variées selon les langues, les algorithmes et les types de caractéristiques discursives.

L'un des objectifs du projet ALVIS est d'intégrer des informations linguistiques dans des moteurs de recherche spécialisés. Dans ce contexte, nous avons conçu une plate-forme d'enrichissement linguistique de documents issus du Web, OGMIOS, exploitant des outils de TAL existants. Les documents peuvent être en français ou en anglais. Cette architecture est distribuée, afin de répondre aux contraintes liées aux traitements de gros volumes de textes, et adaptable, pour permettre l'analyse de sous-langages. La plate-forme est développée en Perl et disponible sous forme de modules CPAN. C'est une structure modulaire dans lequel il est possible d'intégrer de nouvelles ressources ou de nouveaux outils de TAL. On peut ainsi définir des configuration différentes pour différents domaines et types de collections. Cette plateforme robuste permet d'analyser en masse des données issus du web qui sont par essence très hétérogènes. Nous avons évalué les performances de la plateforme sur plusieurs collections de documents. En distribuant les traitements sur vingt machines, une collection de 55 329 documents du domaine de la biologie (106 millions de mots) a été annotée en 35 heures tandis qu'une collection de 48 422 dépêches relatives aux moteurs de recherche (14 millions de mots) a été annotée en 3 heures et 15 minutes.

On observe dans les dictionnaires bilingues une forte asymétrie entre les deux parties d'un même dictionnaire et l'existence de traductions et d'informations « cachées », i.e. pas directement visibles à l'entrée du mot à traduire. Nous proposons une méthodologie de récupération des données cachées ainsi que la « symétrisation » du dictionnaire grâce à un traitement automatique. L'étude d'un certain nombre de verbes et de leurs traductions en plusieurs langues a conduit à l'intégration de toutes les données, visibles ou cachées, au sein d'une base de données unique et multi-lingue. L'exploitation de la base de données a été rendue possible par l'écriture d'un algorithme de création de graphe synonymique qui lie dans un même espace les mots de langues différentes. Le programme qui en découle permettra de générer des dictionnaires paramétrables directement à partir du graphe.

Cet article présente une nouvelle formalisation du modèle de traduction par transfert de la Théorie Sens-Texte. Notre modélisation utilise les grammaires de correspondance polarisées et fait une stricte séparation entre les modèles monolingues, un lexique bilingue minimal et des règles de restructuration universelles, directement associées aux fonctions lexicales syntaxiques.

Dans cet article, nous spécifions les paradigmes de flexion des verbes arabes en respectant la version 9 de LMF (Lexical Markup Framework), future norme ISO 24613 qui traite de la standardisation des bases lexicales. La spécification de ces paradigmes se fonde sur une combinaison des racines et des schèmes. En particulier, nous mettons en relief les terminaisons de racines sensibles aux ajouts de suffixes et ce, afin de couvrir les situations non considérées dans les travaux existants. L'élaboration des paradigmes de flexion verbale que nous proposons est une description en intension d'ArabicLDB (Arabic Lexical DataBase) qui est une base lexicale normalisée pour la langue arabe. Nos travaux sont illustrés par la réalisation d'un conjugueur des verbes arabes à partir d'ArabicLDB.

Nous proposons une nouvelle approche pour l'intégration du TAL dans les systèmes d'apprentissage des langues assisté par ordinateur (ALAO), la stratégie « moinsdisante ». Cette approche tire profit des technologies élémentaires mais fiables du TAL et insiste sur la nécessité de traitements modulaires et déclaratifs afin de faciliter la portabilité et la prise en main didactique des systèmes. Basé sur cette approche, ExoGen est un premier prototype pour la génération automatique d'activités lacunaires ou de lecture d'exemples. Il intègre un module de repérage et de description des réponses des apprenants fondé sur la comparaison entre réponse attendue et réponse donnée. L'analyse des différences graphiques, orthographiques et morpho-syntaxiques permet un diagnostic des erreurs de type fautes d'orthographe, confusions, problèmes d'accord, de conjugaison, etc. La première évaluation d'ExoGen sur un extrait du corpus d'apprenants FRIDA

produit des résultats prometteurs pour le développement de cette approche « moins-disante », et permet d'envisager un modèle d'analyse performant et généralisable à une grande variété d'activités.

Nous présentons une expérience d'extraction automatique des cadres de sous-catégorisation pour 1362 verbes français. Nous exploitons un corpus journalistique richement annoté de 15 000 phrases dont nous extrayons 12 510 occurrences verbales. Nous évaluons dans un premier temps l'extraction des cadres basée sur la fonction des arguments, ce qui nous fournit 39 cadres différents avec une moyenne de 1.54 cadres par lemme. Ensuite, nous adoptons une approche mixte (fonction et catégorie syntaxique) qui nous fournit dans un premier temps 925 cadres différents, avec une moyenne de 3.44 cadres par lemme. Plusieurs méthodes de factorisation, neutralisant en particulier les variantes de réalisation avec le passif ou les pronoms clitiques, sont ensuite appliquées et nous permettent d'aboutir à 235 cadres différents avec une moyenne de 1.94 cadres par verbe. Nous comparons brièvement nos résultats avec les travaux existants pour le français et pour l'anglais.

Nous proposons une formalisation de la décomposition du sens dans le cadre de la Grammaire d'Unification Sens-Texte. Cette formalisation vise une meilleure intégration des décompositions sémantiques dans un modèle global de la langue. Elle repose sur un jeu de saturation de polarités qui permet de contrôler la construction des représentations décomposées ainsi que leur mise en correspondance avec des arbres syntaxiques qui les expriment. Le formalisme proposé est illustré ici dans une perspective de synthèse, mais il s'applique également en analyse.

Les systèmes de questions-réponses (SQR) ont pour but de trouver une information précise extraite d'une grande collection de documents comme le Web. Afin de pouvoir comparer les différentes stratégies possibles pour trouver une telle information, il est important d'évaluer ces systèmes.

L'objectif d'une tâche de validation de réponses est d'estimer si une réponse donnée par un SQR est correcte ou non, en fonction du passage de texte donné comme justification. En 2006, nous avons participé à une tâche de validation de réponses, et dans cet article nous présentons la stratégie que nous avons utilisée. Celle-ci est fondée sur notre propre système de questions-réponses. Le principe est de comparer nos réponses avec les réponses à valider. Nous présentons les résultats obtenus et montrons les extensions possibles. À partir de quelques exemples, nous soulignons les difficultés que pose cette tâche.

Nous voulons traiter des textes chinois automatiquement ; pour ce faire, nous formalisons le vocabulaire chinois, en utilisant principalement des dictionnaires et des grammaires morphologiques et syntaxiques formalisés avec le logiciel NooJ. Nous présentons ici les critères linguistiques qui nous ont permis de construire dictionnaires et grammaires, sachant que l'application envisagée (linguistique de corpus) nous impose certaines contraintes dans la formalisation des unités de la langue, en particulier des composés.

Cet article présente la construction d'un étiqueteur morpho-syntaxique développé pour annoter un corpus de textes kabyles (1 million de mots). Au sein de notre projet, un étiqueteur morpho-syntaxique a été développé et implémenté. Ceci inclut un analyseur morphologique ainsi que l'ensemble de règles de désambiguïsation qui se basent sur l'approche supervisée à base de règles. Pour effectuer le marquage, un jeu d'étiquettes morpho-syntaxiques pour le kabyle est proposé. Les résultats préliminaires sont très encourageants. Nous obtenons un taux d'étiquetage réussi autour de 97 % des textes en prose.

Cet article décrit le traitement automatique du syntagme nominal en grec moderne par le modèle d'analyse syntaxique multi-lingue Fips. L'analyse syntaxique linguistique est focalisée sur les points principaux du DP grec : l'accord entre les constituants fléchis, l'ordre flexible des constituants, la

cliticisation sur les noms et le phénomène de la polydéfinitude. Il est montré comment ces phénomènes sont traités et implémentés dans le cadre de l'analyseur syntaxique FipsGreek, qui met en oeuvre un formalisme inspiré de la grammaire générative chomskyenne.

Le modèle de Gold formalise le processus d'apprentissage d'un langage. Nous présentons dans cet article les avantages et inconvénients de ce cadre théorique contraignant, dans la perspective d'applications en TAL. Nous décrivons brièvement les récentes avancées dans ce domaine, qui soulèvent selon nous certaines questions importantes.

Le présent article décrit deux méthodes d'alignement des propositions : l'une basée sur les méthodes d'appariement des graphes et une autre inspirée de la classification ascendante hiérarchique (CAH). Les deux méthodes sont caractérisées par leur capacité d'alignement des traductions croisées, ce qui était impossible pour beaucoup de méthodes classiques d'alignement des phrases. Contrairement aux résultats obtenus avec l'approche spectrale qui nous paraissent non satisfaisants, l'alignement basé sur la méthode de classification ascendante hiérarchique est prometteur dans la mesure où cette technique supporte bien les traductions croisées.

Cet article propose une approche originale visant la construction d'un lexique sémantique de référence sur le français. Sa principale caractéristique est de pouvoir s'appuyer sur les propriétés morphologiques des lexèmes. La méthode combine en effet des résultats d'analyse morphologique (Namer, 2002;2003), à partir de ressources lexicales de grande taille (nomenclatures du TLF) et des méthodologies d'acquisition d'information lexicale déjà éprouvées (Namer 2005; Sébillot 2002). Le format de représentation choisi, dans le cadre du Lexique Génératif, se distingue par ses propriétés d'expressivité et d'économie. Cette approche permet donc d'envisager la construction d'un lexique de référence sur le français caractérisé par une forte homogénéité tout en garantissant une couverture large, tant du point de vue de la nomenclature que du point de vue des contenus

sémantiques. Une première validation de la méthode fournit une projection quantitative et qualitative des résultats attendus.

Dans cet article, nous présentons les résultats de la campagne d'évaluation EASY des analyseurs syntaxiques du français. EASY a été la toute première campagne d'évaluation comparative des analyseurs syntaxiques du français en mode boîte noire utilisant des mesures objectives quantitatives. EASY fait partie du programme TECHNOLANGUE du Ministère délégué à la Recherche et à l'Éducation, avec le soutien du ministère de délégué à l'industrie et du ministère de la culture et de la communication. Nous exposons tout d'abord la position de la campagne par rapport aux autres projets d'évaluation en analyse syntaxique, puis nous présentos son déroulement, et donnons les résultats des 15 analyseurs participants en fonction des différents types de corpus et des différentes annotations (constituants et relations). Nous proposons ensuite un ensemble de leçons à tirer de cette campagne, en particulier à propos du protocole d'évaluation, de la définition de la segmentation en unités linguistiques, du formalisme et des activités d'annotation, des critères de qualité des données, des annotations et des résultats, et finalement de la notion de référence en analyse syntaxique. Nous concluons en présentant comment les résultats d'EASY se prolongent dans le projet PASSAGE (ANR-06-MDCA-013) qui vient de débuter et dont l'objectif est d'étiqueter un grand corpus par plusieurs analyseurs en les combinant selon des paramètres issus de l'évaluation.

La traduction automatique statistique par séquences de mots est une voie prometteuse. Nous présentons dans cet article deux évolutions complémentaires. La première permet une modélisation de la langue cible dans un espace continu. La seconde intègre des catégories morpho-syntaxiques aux unités manipulées par le modèle de traduction. Ces deux approches sont évaluées sur la tâche Tc-Star. Les résultats les plus intéressants sont obtenus par la combinaison de ces deux méthodes.

Cet article décrit une méthode qui combine des hypothèses graphémiques et phonétiques au niveau de la phrase, à l'aide d'une réprésentation en automates à états finis et d'un modèle de langage, pour la réécriture de phrases tapées au clavier par des dysorthographiques. La particularité des écrits dysorthographiés qui empêche les correcteurs orthographiques d'être efficaces pour cette tâche est une segmentation en mots parfois incorrecte. La réécriture diffère de la correction en ce sens que les phrases réécrites ne sont pas à destination de l'utilisateur mais d'un système automatique, tel qu'un moteur de recherche. De ce fait l'évaluation est conduite sur des versions filtrées et lemmatisées des phrases. Le taux d'erreurs mots moyen passe de 51 % à 20 % avec notre méthode, et est de 0 % sur 43 % des phrases testées.

Le contrôle des hypothèses concurrentes générées par les différents modules qui peuvent intervenir dans des processus de TALN reste un enjeu important malgré de nombreuses avancées en terme de robustesse. Nous présentons dans cet article une méthodologie générique de contrôle exploitant des techniques issues de l'aide multicritère à la décision. À partir de l'ensemble des critères de comparaison disponibles et la formalisation des préférences d'un expert, l'approche proposée évalue la pertinence relative des différents objets linguistiques générés et conduit à la mise en place d'une action de contrôle appropriée telle que le filtrage, le classement, le tri ou la propagation.

Dans cette étude sur le lexique transdisciplinaire des écrits scientifiques, nous souhaitons évaluer dans quelle mesure les méthodes distributionnelles de TAL peuvent faciliter la tâche du linguiste dans le traitement sémantique de ce lexique. Après avoir défini le champ lexical et les corpus exploités, nous testons plusieurs méthodes basées sur des dépendances syntaxiques et observons les proximités sémantiques et les classes établies. L'hypothèse que certaines relations syntaxiques - en particulier les relations de sous-catégorisation ? sont plus appropriées pour établir des classements sémantiques n'apparaît qu'en partie vérifiée. Si les relations de sous-catégorisation génèrent des proximités sémantiques entre les mots de meilleure qualité, cela ne semble pas le cas

pour la classification par voisinage.

LOGUS est un système de compréhension de la langue orale dans le cadre d'un dialogue homme-machine finalisé. Il est la mise en oeuvre d'une approche logique qui utilise différents formalismes afin d'obtenir un système robuste mais néanmoins relativement extensible. Cet article décrit essentiellement l'étape de compréhension en contexte de dialogue implémentée sur LOGUS, développée et testée à partir d'un corpus de réservation hôtelière enregistré et annoté lors des travaux du groupe MEDIA du projet technolangue. Il décrit également les différentes interrogations et conclusions que peut susciter une telle expérience et les résultats obtenus par le système dans la résolution des références. Concernant l'approche elle-même, cette expérience semble montrer que le formalisme adopté pour la représentation sémantique des énoncés est bien adapté à la compréhension en contexte.

Les modèles de Markov cachés (HMM : Hidden Markov Models) (Baum et al., 1970), sont très utilisés en reconnaissance de la parole et depuis quelques années en compréhension de la parole spontanée latine telle que le français ou l'anglais. Dans cet article, nous proposons d'utiliser et d'évaluer la performance de ce type de modèle pour l'interprétation sémantique de la parole arabe spontanée. Les résultats obtenus sont satisfaisants, nous avons atteint un taux d'erreur de l'ordre de 9,9% en employant un HMM à un seul niveau, avec des probabilités tri_grammes de transitions.

Dans cet article nous présentons une évaluation et une analyse des résultats d'une méthode de réordonnancement de réponses pour un système de questions-réponses. Cette méthode propose une sélection des réponses candidates à une question en calculant un coût par transformation d'arbres. Nous présentons une analyse des résultats obtenus sur le corpus Clef 2004-2005 et nos conclusions sur les voies d'amélioration possibles pour notre système.

Le traitement des informations temporelles est crucial pour la compréhension de textes en langue naturelle. Le langage de spécification TimeML a été conçu afin de permettre le repérage et la normalisation des expressions temporelles et des événements dans des textes écrits en anglais. L'objectif des divers projets TimeML a été de formuler un schéma d'annotation pouvant s'appliquer à du texte libre, comme ce que l'on trouve sur le Web, par exemple. Des efforts ont été faits pour l'application de TimeML à d'autres langues que l'anglais, notamment le chinois, le coréen, l'italien, l'espagnol et l'allemand. Pour le français, il y a eu des efforts allant dans ce sens, mais ils sont encore un peu éparpillés. Dans cet article, nous détaillons nos travaux actuels qui visent à élaborer des ressources complètes pour l'annotation de textes en français selon TimeML - notamment un guide d'annotation, un corpus de référence (Gold Standard) et des modules d'annotation automatique.

Nous proposons d'étudier le cas de l'interrogation en Dialogue Homme-Machine au sein d'un système de Question-Réponse à travers le prisme de la Grammaire Interactive. Celle-ci établit un rapport direct entre question et réponse et présuppose que la morphosyntaxe d'une interrogation dépend d'une « réponse escomptée »; l'interlocuteur humain ou machine ayant la possibilité de produire une réponse effective divergente. Nous proposons d'observer la présence des différentes formes de questions dans un corpus issu de l'utilisation du système RITEL. Et nous présentons une expérience menée sur des locuteurs natifs qui nous a permis de mettre en valeur la différence entre réponses effectives produites par nos sujets et réponses présupposées par le contenu intentionnel des questions. Les formalismes ainsi dégagés ont pour but de donner aux systèmes de DHM des fonctionnalités nouvelles comme la capacité à interpréter et à générer de la variabilité dans les énoncés produits.

Dans cet article nous présentons notre démarche pour l'annotation des expressions référentielles désignant les personnes et son utilisation pour la résolution partielle de la référence. Les choix

effectués dans notre implémentation s'inspirent des travaux récents dans le domaine de l'extraction d'information et plus particulièrement de la reconnaissance des entités nommées. Nous utilisons les grammaires locales dans le but d'annoter les entités nommées du type Personne et pour construire, à partir des annotations produites, une base de connaissances extra-linguistiques. Les informations acquises par ce procédé sont ensuite utilisées pour implémenter une méthode de la résolution de la référence pour les syntagmes nominaux coréférentiels.

Nous décrivons la façon dont est formée la phrase japonaise, avec son contenu minimal, la structure des composants d'une phrase simple et l'ordre des mots dans ses composants, les différentes phrases complexes et les possibilités de changements modaux. Le but de cette description est de permettre l'analyse de la phrase japonaise selon des principes universels tout en restant fidèles aux particularités de la langue. L'analyseur syntaxique multilingue FIPS est en cours d'adaptation pour le japonais selon les règles de grammaire qui ont été définies. Bien qu'il fonctionnait alors uniquement pour des langues occidentales, les premiers résultats sont très positifs pour l'analyse des phrases simples, ce qui montre la capacité de Fips à s'adapter à des langues très différentes.

Le projet EmotiRob, financé par l'ANR, a pour but de réaliser un robot compagnon pour des enfants fragilisés. Le projet se décompose en deux sous parties que sont le module de compréhension pour comprendre ce que dit l'enfant et un module d'interaction émotionnelle pour apporter une réponse en simulant des émotions par les mouvements du corps, les traits du visage et par l'émission de petits sons simples. Le module de compréhension dont il est question ici réutilise les travaux du système Logus. La principale difficulté est de faire évoluer le système existant d'un dialogue homme-machine finalisé vers un domaine plus large et de détecter l'état émotionnel de l'enfant. Dans un premier temps, nous présentons le projet EmotiRob et ses spécificités. Ensuite, le système de compréhension de la parole Logus, sur lequel se base ce travail, est présenté en détail. Enfin,

nous présentons les adaptations du système à la nouvelle tâche EmotiRob.

Nous présentons un modèle conceptuel pour la représentation d'opinions, en analysant les éléments qui les composent et quelques propriétés. Ce modèle conceptuel est implémenté et nous en décrivons le jeu d'annotations. Le processus automatique d'annotation de textes en espagnol est effectué par application de règles contextuelles. Un premier sous-ensemble de règles a été écrit pour l'identification de quelques éléments du modèle. Nous analysons les premiers résultats de leur application.

Nous appliquons différents modèles de similarité graphique à la tâche de l'induction de lexiques bilingues entre un dialecte de Suisse allemande et l'allemand standard. Nous comparons des transducteurs stochastiques utilisant des fenêtres glissantes de 1 à 3 caractères, entraînés à l'aide de l'algorithme de maximisation de l'espérance avec des corpus d'entraînement de tailles différentes. Si les transducteurs à uni-grammes donnent des résultats satisfaisants avec des corpus très petits, nous montrons que les transducteurs à bigrammes les dépassent à partir de 750 paires de mots d'entraînement. En général, les modèles entraînés nous ont permis d'améliorer la F-mesure de 7% à 15% par rapport à la distance de Levenshtein.

Dans cet article, nous proposons une approche intégrée localisée pour la génération. Dans cette approche, le traitement intégré des décisions linguistiques est limité à la production des propositions dont les décisions qui concernent leurs générations sont dépendantes. La génération se fait par groupes de propositions de tailles limitées avec traitement intégré des décisions linguistiques qui concernent la production des propositions qui appartiennent au même groupe. Notre approche apporte une solution pour le problème de complexité computationnelle de la génération intégrée classique. Elle fournit ainsi une alternative à la génération séparée (séquentielle ou interactive) qui présente plusieurs défauts mais qui est implémentée de manière répandue dans les systèmes de

générations existants.

L'objectif de cet article est la présentation d'un système de génération automatique de dictionnaires électroniques de la langue arabe classique, développé au sein de laboratoire UTIC (unité de Monastir). Dans cet article, nous présenterons, les différentes étapes de réalisation, et notamment la génération automatique de ces dictionnaires se basant sur une théorie originale : les Conditions de Structures Morphomatiques (CSM), et les matrices lexicales. Ce système rentre dans le cadre des deux projets MIRTO et OREILLODULE réalisés dans les deux laboratoires LIDILEM de Grenoble et UTIC Monastir de Tunisie

Dans le contexte de la détection de plagiats, le repérage de citations et de ses constituants est primordial puisqu'il peut aider à évaluer le caractère licite ou illicite d'une reprise (source citée ou non). Nous proposons ici une étude quantitative et qualitative des citations extraites d'un corpus que nous avons auparavant construit. Cette étude a pour but de tracer des axes de recherche vers une méthode de repérage automatique des citations.

Nous présentons des travaux réalisés dans le domaine des systèmes de questions réponses (SQR) utilisant des questions enchainées. La recherche des documents dans un SQR est perturbée par l'absence d'informations sur la valeur à accorder aux éléments de texte éventuellement utiles à la recherche d'informations qui figurent dans les questions liées. Les récentes campagnes d'évaluation montrent que ce problème est sous-estimé, et n'a pas fait l'oeuvre de technique dédiée. Afin d'améliorer la recherche des documents dans un SQR nous étudions une nouvelle méthode pour organiser les informations liées aux interactions entre questions. Celle-ci se base sur l'exploitation d'une structure de données adaptée à la transmission des informations des questions liées jusqu'au moteur d'interrogation.

La correction grammaticale automatique du français est une fonctionnalité qui fait cruellement défaut à la communauté des utilisateurs de logiciels libres. Dans le but de combler cette lacune, nous avons travaillé à l'adaptation au français d'un outil initialement développé pour une langue étrangère. Ce travail nous a permis de montrer que les approches classiques du traitement automatique des langues utilisées dans le domaine ne sont pas appropriées. Pour y remédier, nous proposons de faire évoluer les formalismes des correcteurs en intégrant les principes linguistiques de la segmentation en chunks et de l'unification. Bien qu'efficace, cette évolution n'est pas suffisante pour obtenir un bon correcteur grammatical du français. Nous envisageons alors une nouvelle approche de la problématique.

Cet article présente un projet de repérage, d'extraction et d'annotation d'informations temporelles, d'informations spatiales et d'objets touristiques dans des pages Web afin d'alimenter la base de connaissance d'un portail touristique. Nous portons une attention particulière aux différences qui distinguent le repérage d'information dans des pages Web du repérage d'informations dans des documents structurés. Après avoir introduit et classifié les différentes informations à extraire, nous nous intéressons à la façon de lier ces informations entre elles (par exemple apparier une information d'ouverture et un restaurant) et de les annoter. Nous présentons également le logiciel que nous avons réalisé afin d'effectuer cette opération d'annotation ainsi que les premiers résultats obtenus. Enfin, nous nous intéressons aux autres types de marques que l'on trouve dans les pages Web, les marques sémiotiques en particulier, dont l'analyse peut être utile à l'interprétation des pages.

La présente contribution part de nos constats réalisés à partir des résultats d'évaluation de notre système d'alignement des propositions de textes français-japonais. La présence importante de structures fondamentalement difficiles à aligner et les résultats peu satisfaisants de différentes méthodes de mise en correspondance des mots nous ont finalement amenés à remettre en cause

l'existence même d'équivalence au niveau des propositions syntaxiques entre le français et le japonais. Afin de compenser les défauts que nous avons découverts, nous proposons des opérations permettant de restaurer l'équivalence des propositions alignées et d'améliorer la qualité des corpus alignés.

Nous définissons un formalisme, les grammaires rationnelles d'arbres avec traits, et une traduction des grammaires d'arbres adjoints avec traits vers ce nouveau formalisme. Cette traduction préserve les structures de dérivation de la grammaire d'origine en tenant compte de l'unification de traits. La construction peut être appliquée aux réalisateurs de surface qui se fondent sur les arbres de dérivation.

Cet article présente une méthode basée sur des calculs de distance et une analyse sémantique et syntaxique pour la segmentation thématique de texte. Pour évaluer cette méthode nous la comparons à un un algorithme lexical très connu : c99. Nous testons les deux méthodes sur un corpus de discours politique français et comparons les résultats. Les deux conclusions qui ressortent de notre expérience sont que les approches sont complémentaires et que les protocoles d'évaluation actuels sont inadaptés.

Cet article propose un modèle de langage dédié au dialogue homme-machine, ainsi que des algorithmes d'analyse et de génération. L'originalité de notre approche est de faire reposer l'analyse et la génération sur les mêmes connaissances, essentiellement sémantiques. Celles-ci sont structurées sous la forme d'une bibliothèque de concepts, et de formes d'usage associées aux concepts. Les algorithmes, quant à eux, sont fondés sur un double principe de correspondance entre des offres et des attentes, et d'un calcul heuristique de score.

Ces travaux présentent une extension des représentations formelles pour la sémantique, de l'outil

de traitement automatique des langues de Orange Labs 1. Nous abordons ici uniquement des questions relatives à la construction des représentations sémantiques, dans le cadre de l'analyse linguistique. Afin d'obtenir des représentations plus fines de la structure argumentale des énoncés, nous incluons des concepts issus de la DRT dans le système de représentation basé sur les graphes sémantiques afin de rendre compte de la notion de portée.

La production de lexiques est une activité indispensable mais complexe, qui nécessite, quelle que soit la méthode de création utilisée (acquisition automatique ou manuelle), une validation humaine. Nous proposons dans ce but une plate-forme Web librement disponible, appelée Sylva (Systematic lexicon validator). Cette plate-forme a pour caractéristiques principales de permettre une validation multi-niveaux (par des validateurs, puis un expert) et une traçabilité de la ressource. La tâche de l'expert(e) linguiste en est allégée puisqu'il ne lui reste à considérer que les données sur lesquelles il n'y a pas d'accord inter-validateurs.

La croissance exponentielle de l'Internet a permis le développement de sites d'offres d'emploi en ligne. Le système E-Gen (Traitement automatique d'offres d'emploi) a pour but de permettre l'analyse et la catégorisation d'offres d'emploi ainsi qu'une analyse et classification des réponses des candidats (Lettre de motivation et CV). Nous présentons les travaux réalisés afin de résoudre la seconde partie : on utilise une représentation vectorielle de texte pour effectuer une classification des pièces jointes contenus dans le mail à l'aide de SVM. Par la suite, une évaluation de la candidature est effectuée à l'aide de différents classifieurs (SVM et n-grammes de mots).

Dans cet article, nous présentons un nouveau langage permettant d'écrire des relations rationnelles compilées en automates finis. Les deux caractéristiques innovantes de ce langage sont de pourvoir décrire des relations à plusieurs niveaux, pas nécessairement deux et d'utiliser diverses unités d'analyse pour exprimer les liens entre niveaux. Cela permet d'aligner de façon fine des

représentations multiples.

L'évaluation pour le dialogue homme-machine ne se caractérise pas par l'efficacité, l'objectivité et le consensus que l'on observe dans d'autres domaines du traitement automatique des langues. Les systèmes de dialogue oraux et multimodaux restent cant-onnés à des domaines applicatifs restreints, ce qui rend difficiles les évaluations comparatives ou normées. De plus, les avancées technologiques constantes rendent vite obsolètes les paradigmes d'évaluation et ont pour conséquence une multiplication de ceux-ci. Des solutions restent ainsi à trouver pour améliorer les méthodes existantes et permettre des diagnostics plus automatisés des systèmes. Cet article se veut un ensemble de réflexions autour de l'évaluation de la multimodalité dans les systèmes à forte composante linguistique. Des extensions des paradigmes existants sont proposées, en particulier DQR/DCR, sachant que certains sont mieux adaptés que d'autres au dialogue multimodal. Des conclusions et perspectives sont tirées sur l'avenir de l'évaluation pour le dialogue homme-machine.

Cet article présente POLYMOTS, une base de données lexicale contenant huit mille mots communs en français. L'originalité de l'approche proposée tient à l'analyse des mots. En effet, à la différence d'autres bases lexicales représentant la morphologie dérivationnelle des mots à partir d'affixes, ici l'idée a été d'isoler un radical commun à un ensemble de mots d'une même famille. Nous avons donc analysé les formes des mots et, par comparaison phonologique (forme phonique comparable) et morphologique (continuité de sens), nous avons regroupé les mots par familles, selon le type de radical phonologique. L'article présente les fonctionnalités de la base et inclut une discussion sur les applications et les perspectives d'une telle ressource.

La génération de néologismes construits pose des problèmes dans un système de traduction automatique, notamment au moment de la sélection du préfixe dans les formations préfixées, quand certains préfixes paraissent pouvoir alterner. Nous proposons une étude « extensive », qui vise à

rechercher dans de larges ressources textuelles (l'Internet) des formes préfixées générées automatiquement, dans le but d'individualiser les paramètres qui favorisent l'un des préfixes ou qui, au contraire, permettent cette alternance. La volat-ilité de cette ressource textuelle nécessite certaines précautions dans la méthodologie de décompte des données extraites.

Cet article présente l'utilisation de la plate-forme CasSys pour la segmentation de la parole conversationnelle (chunking) à l'aide de cascades de transducteurs Unitex. Le système que nous présentons est utilisé dans le cadre du projet ANR EPAC. Ce projet a pour objectif l'indexation et l'annotation automatique de grands flux de parole issus d'émissions télévisées ou radiophoniques. Cet article présente tout d'abord l'adaptation à ce type de données d'un système antérieur de chunking (Romus) qui avait été développé pour le dialogue oral homme-machine. Il décrit ensuite les principaux problèmes qui se posent à l'analyse : traitement des disfluences de l'oral spontané, mais également gestion des erreurs dues aux étapes antérieures de reconnaissance de la parole et d'étiquetage morpho-syntaxique.

Cet article porte sur le regroupement automatique de documents sur une base événementielle. Après avoir précisé la notion d'événement, nous nous intéressons à la représentation des documents d'un corpus de dépêches, puis à une approche d'apprentissage pour réaliser les regroupements de manière non supervisée fondée sur k-means. Enfin, nous évaluons le système de regroupement de documents sur un corpus de taille réduite et nous discutons de l'évaluation quantitative de ce type de tâche.

Nous évaluons le recours à des techniques de traduction à base de segments syntaxiquement motivés, seules ou en combinaison avec des techniques à base de segments non motivés, et nous comparons les apports respectifs de l'analyse en constituants et de l'analyse en dépendances dans ce cadre. À partir d'un corpus parallèle Anglais?Français, nous construisons automatiquement deux

corpus d'entraînement arborés, en constituants et en dépendances, alignés au niveau sous-phrastique et en extrayons des correspondances bilingues entre mots et syntagmes motivées syntaxiquement. Nous mesurons automatiquement la qualité de la traduction obtenue par un système à base de segments. Les résultats montrent que la combinaison des correspondances bilingues non motivées et motivées sur le plan syntaxique améliore la qualité de la traduction quel que soit le type d'analyse considéré. Par ailleurs, le gain en qualité est plus important avec le recours à l'analyse en dépendances au regard des constituants.

Dans le contexte de la recherche de plagiat, le repérage de citations et de ses constituants est primordial puisqu'il peut amener à évaluer le caractère licite ou illicite d'une reprise (source citée ou non). Nous proposons ici une comparaison de méthodes automatiques pour le repérage de ces informations et rapportons une évaluation quantitative de celles-ci. Un corpus d'écrits journalistiques français a été manuellement annoté pour nous servir de base d'apprentissage et de test.

Il peut être difficile d'attribuer une seule valeur illocutoire à un énoncé dans un dialogue. En premier lieu, un énoncé peut comporter plusieurs segments de discours ayant chacun leur valeur illocutoire spécifique. De plus, un seul segment peut s'analyser en tant qu'acte de langage composite, regroupant par exemple la formulation d'une question et l'émission simultanée d'une information. Enfin, la structure du dialogue en termes d'échanges et de séquences peut être déterminante dans l'identification de l'acte, et peut également apporter une valeur illocutoire supplémentaire, comme celle de clore la séquence en cours. Dans le but de déterminer la réaction face à un tel acte de dialogue composite, nous présentons une approche théorique pour l'analyse des actes de dialogue en fonction du contexte de tâche et des connaissances des interlocuteurs. Nous illustrons sur un exemple nos choix de segmentation et d'identification des actes composites, et nous présentons les grandes lignes d'une stratégie pour déterminer la réaction qui semble être la plus pertinente.

La représentation évènementielle des déplacements de personnes dans des dépêches épidémiologiques est d'une grande importance pour une compréhension détaillée du sens de ces dépêches. La dissémination des composants d'une telle représentation dans les dépêches rend difficile l'accès à leurs contenus. Ce papier décrit un système d'extraction d'information utilisant des cascades de transducteurs à nombre d'états fini qui ont permis la réalisation de trois tâches : la reconnaissance des entités nommées, l'annotation et la représentation des composants ainsi que la représentation des structures évènementielles. Nous avons obtenu une moyenne de rappel de 80,93% pour la reconnaissance des entités nommées et de 97,88% pour la représentation des composants. Ensuite, nous avons effectué un travail de normalisation de cette représentation par la résolution de certaines anaphores pronominales. Nous avons obtenu une valeur moyenne de précision de 81,72% pour cette résolution.

L'augmentation rapide des échanges et des communications pluriculturels, en particulier sur internet, intensifie les besoins d'outils multi-lingues y compris de traduction. Cet article décrit un projet en cours au LATL pour le développement d'un système de traduction multi-lingue basé sur un modèle linguistique abstrait et largement générique, ainsi que sur un modèle logiciel basé sur la notion d'objet. Les langues envisagées dans la première phase de ce projet sont l'allemand, le français, l'italien, l'espagnol et l'anglais.

L'appariement d'entités nommées consiste à regrouper les différentes formes sous lesquelles apparaît une entité. Pour cela, des mesures de similarité textuelle sont généralement utilisées. Nous proposons de combiner plusieurs mesures afin d'améliorer les performances de la tâche d'appariement. À l'aide d'expériences menées sur deux corpus, nous montrons la pertinence de l'apprentissage supervisé dans ce but, particulièrement avec l'algorithme C4.5.

Nous discutons du sens des graphes sémantiques, notamment de ceux utilisés en Théorie

Sens-Texte. Nous leur donnons un sens précis, éventuellement sous-spécifié, grâce à une traduction simple vers une formule de Minimal Recursion Se-mantics qui couvre les cas de prédications multiples sur plusieurs entités, de prédication d'ordre supérieur et de modalités.

Cet article présente un formalisme de représentation des connaissances qui a été utilisé pour fournir des annotations sémantiques de haut niveau pour le corpus de dialogue oral MEDIA. Ces annotations en structures sémantiques, basées sur le paradigme FrameNet, sont obtenues de manière incrémentale et partiellement automatisée. Nous décrivons le processus d'interprétation automatique qui permet d'obtenir des compositions sémantiques et de générer des hypothèses de frames par inférence. Le corpus MEDIA est un corpus de dialogues en langue française dont les tours de parole de l'utilisateur ont été manuellement transcrits et annotés (niveaux mots et constituants sémantiques de base). Le processus proposé utilise ces niveaux pour produire une annotation de haut niveau en frames sémantiques. La base de connaissances développée (définitions des frames et règles de composition) est présentée, ainsi que les résultats de l'annotation automatique.

Les moteurs de recherches sur le web produisent des résultats comparables et assez satisfaisants pour la recherche de documents écrits en caractères latins. Cependant, ils présentent de sérieuses lacunes dès que l'ont s'intéresse à des langues peu dotées ou des langues sémitiques comme l'arabe. Dans cet article nous présentons une étude analytique et qualitative de la recherche d'information en langue arabe en mettant l'accent sur l'insuffisance des outils de recherche actuels, souvent mal adaptés aux spécificités de la langue arabe. Pour argumenter notre analyse, nous présentons des résultats issus d'observations et de tests autour de certains phénomènes linguistiques de l'arabe écrit. Pour la validation des ces observations, nous avons testé essentiellement le moteur de recherche Google.

Nous présentons un système de normalisation de la variation syntaxique qui permet de mieux reconnaître la relation d'implication textuelle entre deux phrases. Le système est évalué sur une suite de tests comportant 2 520 paires test et les résultats montrent un gain en précision par rapport à un système de base variant entre 29.8 et 7-8.5 points la complexité des cas considérés.

Cet article aborde l'étude des expressions temporelles qui font référence directement à des unités de temps relatives aux divisions courantes des calendriers, que nous qualifions d'expressions calendaires (EC). Nous proposons une modélisation de ces expressions en définissant une algèbre d'opérateurs qui sont liés aux classes de marqueurs linguistiques qui apparaissent dans les EC. A partir de notre modélisation, une vue calendaire est construite dans la plate-forme de visualisation et navigation textuelle NaviTexte, visant le support à la lecture de textes. Enfin, nous concluons sur les perspectives offertes par le développement d'une première application de navigation temporelle.

Dans cet article, nous proposons une méthode pour identifier, dans un texte en français, l'ensemble des expressions adverbiales de localisation temporelle, ainsi que tous les verbes, noms et adjectifs dénotant une éventualité (événement ou état). Cette méthode, en plus d'identifier ces expressions, extrait certaines informations sémantiques : la valeur de la localisation temporelle selon la norme TimeML et le type des éventualités. Pour les expressions adverbiales de localisation temporelle, nous utilisons une cascade d'automates, alors que pour l'identification des événements et états nous avons recours à une analyse complète de la phrase. Nos résultats sont proches de travaux comparables sur l'anglais, en l'absence d'évaluation quantitative similaire sur le français.

Nous présentons une méthode de segmentation de journaux radiophoniques en sujets, basée sur la prise en compte d'indices lexicaux, syntaxiques et acoustiques. Partant d'un modèle statistique existant de segmentation thématique, exploitant la notion de cohésion lexicale, nous étendons le formalisme pour y inclure des informations d'ordre syntaxique et acoustique. Les résultats

expérimentaux montrent que le seul modèle de cohésion lexicale ne suffit pas pour le type de documents étudié en raison de la taille variable des segments et de l'absence d'un lien direct entre segment et thème. L'utilisation d'informations syntaxiques et acoustiques permet une amélioration substantielle de la segmentation obtenue.

Nous présentons dans cet article une méthode d'extraction automatique d'informations sur des textes de très petite taille, faiblement structurés. Nous travaillons sur des textes dont la rédaction n'est pas normalisée, avec très peu de mots pour caractériser chaque information. Les textes ne contiennent pas ou très peu de phrases. Il s'agit le plus souvent de morceaux de phrases ou d'expressions composées de quelques mots. Nous comparons plusieurs méthodes d'extraction, dont certaines sont entièrement automatiques. D'autres utilisent en partie une connaissance du domaine que nous voulons réduite au minimum, de façon à minimiser le travail manuel en amont. Enfin, nous présentons nos résultats qui dépassent ce dont il est fait état dans la littérature, avec une précision équivalente et un rappel supérieur.

Dans cet article, nous présentons une discussion sur la combinaison de différents scores et critères numériques pour la sélection finale d'une réponse dans la partie en charge des questions factuelles du système de Questions/Réponses développé au LIA. Ces scores et critères numériques sont dérivés de ceux obtenus en sortie de deux composants cruciaux pour notre système : celui de sélection des passages susceptibles de contenir une réponse et celui d'extraction et de sélection d'une réponse. Ils sont étudiés au regard de leur expressivité. Des comparaisons sont faites avec des approches de sélection de passages mettant en oeuvre des scores conventionnels en recherche d'information. Parallèlement, l'influence de la taille des contextes (en nombre de phrases) est évaluée. Cela permet de mettre en évidence que le choix de passages constitués de trois phrases autour d'une réponse candidate, avec une sélection des réponses basée sur une combinaison entre un score de passage de type Lucene ou Cosine et d'un score de compacité

apparaît comme un compromis intéressant.

Les connecteurs discursifs ont on rôle important dans l'interprétation des discours (dialogiques ou pas), donc lorsqu'il s'agit de produire des énoncés, le choix des mots qui relient les énoncés (par exemple, en dialogue oral) s'avère essentiel pour assurer la compréhension des visées illocutoires des locuteurs. En linguistique computationnelle, le problème a été abordé surtout au niveau de l'interprétation des discours monologiques, tandis que pour le dialogue, les recherches se sont limitées en général à établir une correspondance quasiment biunivoque entre relations rhétoriques et connecteurs. Dans ce papier nous proposons un mécanisme pour guider la génération des connecteurs concessifs en dialogue, à la fois du point de vue discursif et sémantique ; chaque connecteur considéré sera contraint par un ensemble de conditions qui prennent en compte la cohérence du discours et la pertinence sémantique de chaque mot concerné. Les contraintes discursives, exprimées dans un formalisme dérivé de la SDRT (« Segmented Discourse Representation Theory ») seront plongées dans des contraintes sémantiques sur les connecteurs, proposées par l'école genevoise (Moeschler), pour enfin évaluer la cohérence du discours résultant de l'emploi de ces connecteurs.

Cet article présente une modélisation du principe d'ancrage (grounding) pour la robustesse des systèmes de dialogue finalisés. Ce principe, décrit dans (Clark & Schaefer, 1989), suggère que les participants à un dialogue fournissent des preuves de compréhension afin d'atteindre la compréhension mutuelle. Nous explicitons une définition computationnelle du principe d'ancrage fondée sur des jugements de compréhension qui, contrairement à d'autres modèles, conserve une motivation pour l'expression de la compréhension. Nous déroulons enfin le processus d'ancrage sur un exemple tiré de l'implémentation du modèle.

Dans cet article, nous présentons des applications du système Enertex au Traitement Automatique

de la Langue Naturelle. Enertex est basé sur l'énergie textuelle, une approche par réseaux de neurones inspirée de la physique statistique des systèmes magnétiques. Nous avons appliqué cette approche aux problèmes du résumé automatique multi-documents et de la détection de frontières thématiques. Les résultats, en trois langues : anglais, espagnol et français, sont très encourageants.

Dans cet article, nous présentons les améliorations que nous avons apportées au système ExtraNews de résumé automatique de documents multiples. Ce système se base sur l'utilisation d'un algorithme génétique qui permet de combiner les phrases des documents sources pour former les extraits, qui seront croisés et mutés pour générer de nouveaux extraits. La multiplicité des critères de sélection d'extraits nous a inspiré une première amélioration qui consiste à utiliser une technique d'optimisation multi-objectif en vue d'évaluer ces extraits. La deuxième amélioration consiste à intégrer une étape de pré-filtrage de phrases qui a pour objectif la réduction du nombre des phrases des textes sources en entrée. Une évaluation des améliorations apportées à notre système est réalisée sur les corpus de DUC'04 et DUC'07.

Dans cette étude, nous nous intéressons à des algorithmes de recherche locale pour la traduction statistique à base de segments (phrase-based machine translation). Les algorithmes que nous étudions s'appuient sur une formulation complète d'un état dans l'espace de recherche contrairement aux décodeurs couramment utilisés qui explorent l'espace des préfixes des traductions possibles. Nous montrons que la recherche locale seule, permet de produire des traductions proches en qualité de celles fournies par les décodeurs usuels, en un temps nettement inférieur et à un coût mémoire constant. Nous montrons également sur plusieurs directions de traduction qu'elle permet d'améliorer de manière significative les traductions produites par le système à l'état de l'art Pharaoh (Koehn, 2004).

Cet article présente une architecture inspirée des systèmes de reconnaissance vocale pour

effectuer une normalisation orthographique de messages en « langage SMS ». Nous décrivons notre système de base, ainsi que diverses évolutions de ce système, qui permettent d'améliorer sensiblement la qualité des normalisations produites.

Cet article étudie la relation entre les grammaires d'arbres adjoints à composantes multiples avec tuples d'arbres (TT-MCTAG), un formalisme utilisé en linguistique informatique, et les grammaires à concaténation d'intervalles (RCG). Les RCGs sont connues pour décrire exactement la classe PTIME, il a en outre été démontré que les RCGs « simples » sont même équivalentes aux systèmes de réécriture hors-contextes linéaires (LCFRS), en d'autres termes, elles sont légèrement sensibles au contexte. TT-MCTAG a été proposé pour modéliser les langages à ordre des mots libre. En général ces langages sont NP-complets. Dans cet article, nous définissons une contrainte additionnelle sur les dérivations autorisées par le formalisme TT-MCTAG. Nous montrons ensuite comment cette forme restreinte de TT-MCTAG peut être convertie en une RCG simple équivalente. Le résultat est intéressant pour des raisons théoriques (puisqu'il montre que la forme restreinte de TT-MCTAG est légèrement sensible au contexte), mais également pour des raisons pratiques (la transformation proposée ici a été utilisée pour implanter un analyseur pour TT-MCTAG).

Cet article décrit un analyseur syntaxique pour grammaires de dépendance lexicalisées. Le formalisme syntaxique se caractérise par une factorisation des contraintes syntaxiques qui se manifeste dans la séparation entre dépendance et ordre linéaire, la spécification fonctionnelle (plutôt que syntagmatique) des dépendants, la distinction entre dépendants valenciels (la sous-catégorisation) et non valenciels (les circonstants) et la saturation progressive des arbres. Ceci résulte en une formulation concise de la grammaire à un niveau très abstrait et l'élimination de la reduplication redondante des informations due aux réalisations alternatives des dépendants ou à leur ordre. Les arbres élémentaires (obtenus à partir des formes dans l'entrée) et dérivés sont combinés entre eux par adjonction d'un arbre dépendant saturé à un arbre régissant, moyennant

l'unification des noeuds et des relations. La dérivation est réalisée grâce à un analyseur chart bi-directionnel.

Pour la formalisation du lexique et de la grammaire de dialectes étroitement apparentés, il peut se révéler utile de factoriser une partie du travail de modélisation. Les soussystèmes linguistiques isomorphes dans les différents dialectes peuvent alors faire l'objet d'une description commune, les différences étant spécifiées par ailleurs. Cette démarche aboutit à un modèle de grammaire à couches : le noyau est commun à la famille de dialectes, et une couche superficielle détermine les caractéristiques de chacun. Nous appliquons ce procédé à la famille des langues créoles à base lexicale française de l'aire américano-caraïbe.

Nous montrons qu'il est possible d'obtenir une analyse syntaxique statistique satisfaisante pour le français sur du corpus journalistique, à partir des données issues du French Treebank du laboratoire LLF, à l'aide d'un algorithme d'analyse non lexicalisé.

Cet article décrit la construction d'un Wordnet Libre du Français (WOLF) à partir du Princet-on Word-Net et de diverses ressources multi-lingues. Les lexèmes polysémiques ont été traités au moyen d'une approche reposant sur l'alignement en mots d'un corpus parallèle en cinq langues. Le lexique multi-lingue extrait a été désambiguïsé sémantiquement à l'aide des wordnets des langues concernées. Par ailleurs, une approche bilingue a été suffisante pour construire de nouvelles entrées à partir des lexèmes monosémiques. Nous avons pour cela extrait des lexiques bilingues à partir deWikipédia et de thésaurus. Le wordnet obtenu a été évalué par rapport au wordnet français issu du projet EuroWord-Net. Les résultats sont encourageants, et des applications sont d'ores et déjà envisagées.

Les informations lexicales, indispensables pour les tâches réalisées en TALN, sont difficiles à

collecter. En effet, effectuée manuellement, cette tâche nécessite la compétence d'experts et la durée nécessaire peut être prohibitive, alors que réalisée automatiquement, les résultats peuvent être biaisés par les corpus de textes retenus. L'approche présentée ici consiste à faire participer un grand nombre de personnes à un projet contributif en leur proposant une application ludique accessible sur le web. A partir d'une base de termes préexistante, ce sont ainsi les joueurs qui vont construire le réseau lexical, en fournissant des associations qui ne sont validées que si elles sont proposées par au moins une paire d'utilisateurs. De plus, ces relations typées sont pondérées en fonction du nombre de paires d'utilisateurs qui les ont proposées. Enfin, nous abordons la question de la détermination des différents sens d'usage d'un terme, en analysant les relations entre ce terme et ses voisins immédiats dans le réseau lexical, avant de présenter brièvement la réalisation et les premiers résultats obtenus.

Le présent papier s'intéresse à l'élaboration des dictionnaires électroniques arabes à usage éditorial. Il propose un modèle unifié et normalisé de ces dictionnaires en se référant à la future norme LMF (Lexical Markup Framework) ISO 24613. Ce modèle permet de construire des dictionnaires extensibles, sur lesquels on peut réaliser, grâce à une structuration fine et standard, des fonctions de consultation génériques adaptées aux besoins des utilisateurs. La mise en oeuvre du modèle proposé est testée sur des dictionnaires existants de la langue arabe en utilisant, pour la consultation, le système ADIQTO (Arabic DIctionary Query TOols) que nous avons développé pour l'interrogation générique des dictionnaires normalisés de l'arabe.

Cette étude propose une analyse et une modélisation des relations de polysémie dans le lexique électronique anglais Word-Net. Elle exploite pour cela la hiérarchie des concepts (représentés par des synsets), et la définition associée à chacun de ces concepts. Le résultat est constitué d'un ensemble de règles qui nous ont permis d'identifier d'une façon largement automatisée, avec une précision voisine de 91%, plus de 2100 paires de synsets liés par une relation de polysémie

régulière. Notre méthode permet aussi une désambiguïsation lexicale partielle des mots de la définition associée à ces synsets.

Dans cet article, nous présentons une nouvelle approche pour la traduction automatique fondée sur les triggers inter-langues. Dans un premier temps, nous expliquons le concept de triggers inter-langues ainsi que la façon dont ils sont déterminés. Nous présentons ensuite les différentes expérimentations qui ont été menées à partir de ces triggers afin de les intégrer au mieux dans un processus complet de traduction automatique. Pour cela, nous construisons à partir des triggers inter-langues des tables de traduction suivant différentes méthodes. Nous comparons par la suite notre système de traduction fondé sur les triggers inter-langues à un système état de l'art reposant sur le modèle 3 d'IBM (Brown eta al., 1993). Les tests menés ont montré que les traductions automatiques générées par notre système améliorent le score BLEU (Papineni & al., 2001) de 2, 4% comparé à celles produites par le système état de l'art.

Cet article présente une approche pour obtenir des paraphrases pour de courts segments de texte qui peuvent aider un rédacteur à reformuler localement des textes. La ressource principale utilisée est une table d'alignements bilingues de segments d'un système de traduction automatique statistique. Un segment marqué par le rédacteur est tout d'abord traduit dans une langue pivot avant d'être traduit à nouveau dans la langue d'origine, ce qui est permis par la nature même de la ressource bilingue utilisée sans avoir recours à un processus de traduction complet. Le cadre proposé permet l'intégration et la combinaison de différents modèles d'estimation de la qualité des paraphrases. Des modèles linguistiques tentant de prendre en compte des caractéristiques des paraphrases de courts segments de textes sont proposés, et une évaluation est décrite et ses résultats analysés. Les domaines d'application possibles incluent, outre l'aide à la reformulation, le résumé et la réécriture des textes pour répondre à des conventions ou à des préférences stylistiques. L'approche est critiquée et des perspectives d'amélioration sont proposées.

Contrairement à une idée répandue, les architectures linguistiques et computationnelles des systèmes de traduction automatique sont indépendantes. Les premières concernent le choix des représentations intermédiaires, les secondes le type d'algorithme, de programmation et de ressources utilisés. Il est ainsi possible d'utiliser des méthodes de calcul « expertes » ou « empiriques » pour construire diverses phases ou modules de systèmes d'architectures linguistiques variées. Nous terminons en donnant quelques éléments pour le choix de ces architectures en fonction des situations traductionnelles et des ressources disponibles, en termes de dictionnaires, de corpus, et de compétences humaines.

Dans cet article, nous proposons une méthode visant à corriger et à associer dynamiquement de nouveaux types sémantiques dans le cadre de systèmes de détection automatique d'entités nommées (EN). Après la détection des entités nommées et aussi de manière plus générale des noms propres dans les textes, une vérification de compatibilité de types sémantiques est effectuée non seulement pour confirmer ou corriger les résultats obtenus par le système de détection d'EN, mais aussi pour associer de nouveaux types non couverts par le système de détection d'EN. Cette vérification est effectuée en utilisant l'information syntaxique associée aux EN par un système d'analyse syntaxique robuste et en confrontant ces résultats avec la ressource sémantique Word-Net. Les résultats du système de détection d'EN sont alors considérablement enrichis, ainsi que les étiquettes sémantiques associées aux EN, ce qui est particulièrement utile pour l'adaptation de systèmes de détection d'EN à de nouveaux domaines.

Nous présentons une méthode non supervisée de désambiguïsation d'entités nommées, basée sur l'exploitation des treillis de Galois. Nous réalisons une analyse de concepts formels à partir de relations entre des entités nommées et leurs contextes syntaxiques extraits d'un corpus d'apprentissage. Le treillis de Galois résultant fournit des concepts qui sont utilisés comme des

étiquettes pour annoter les entités nommées et leurs contextes dans un corpus de test. Une évaluation en cascade montre qu'un système d'apprentissage supervisé améliore la classification des entités nommées lorsqu'il s'appuie sur l'annotation réalisée par notre système de désambiguïsation non supervisée.

Dans cet article, nous décrivons la méthode que nous avons développée pour la résolution de métonymie des entités nommées dans le cadre de la compétition SemEval 2007. Afin de résoudre les métonymies sur les noms de lieux et noms d'organisation, tel que requis pour cette tâche, nous avons mis au point un système hybride basé sur l'utilisation d'un analyseur syntaxique robuste combiné avec une méthode d'analyse distributionnelle. Nous décrivons cette méthode ainsi que les résultats obtenus par le système dans le cadre de la compétition SemEval 2007.

Cet article se propose d'étudier les relations sémantiques reliant base et expansion au sein des termes médicaux arabes de type « N+N », particulièrement ceux dont la base est un déverbal. En étudiant les relations sémantiques établies par une base déverbale, ce travail tente d'attirer l'attention sur l'interpénétration du sémantique et du morpho-syntaxique ; il montre que, dans une large mesure, la structure morpho-syntaxique de la base détermine l'éventail des possibilités relationnelles. La découverte de régularités dans le comportement de la base déverbale permet de prédire le type de relations que peut établir cette base avec son expansion pavant ainsi la voie à un traitement automatique et un travail d'étiquetage sémantique des textes médicaux arabes.

Nous montrons dans cet article qu'il existe une corrélation étroite existant entre la qualité de l'étiquetage morpho-syntaxique et les performances des chunkers. Cette corrélation devient linéaire lorsque la taille des chunks est limitée. Nous appuyons notre démonstration sur la base d'une expérimentation conduite suite à la campagne d'évaluation Passage 2007 (de la Clergerie et al., 2008). Nous analysons pour cela les comportements de deux analyseurs ayant participé à cette

campagne. L'interprétation des résultats montre que la tâche de chunking, lorsqu'elle vise des chunks courts, peut être assimilée à une tâche de "super-étiquetage".

L'indexation est une composante importante de tout système de recherche d'information. Dans MEDLINE, la base documentaire de référence pour la littérature du domaine biomédical, le contenu des articles référencés est indexé à l'aide de descripteurs issus du thésaurus MeSH. Avec l'augmentation constante de publications à indexer pour maintenir la base à jour, le besoin d'outils automatiques se fait pressant pour les indexeurs. Dans cet article, nous décrivons l'utilisation et l'adaptation de la Programmation Logique Inductive (PLI) pour découvrir des règles d'indexation permettant de générer automatiquement des recommandations d'indexation pour MEDLINE. Les résultats obtenus par cette approche originale sont très satisfaisants comparés à ceux obtenus à l'aide de règles manuelles lorsque celles-ci existent. Ainsi, les jeux de règles obtenus par PLI devraient être prochainement intégrés au système produisant les recommandations d'indexation automatique pour MEDLINE.

Dans le domaine de la classification supervisée et semi-supervisée, cet article présente un contexte favorable à l'application de méthodes statistiques de classification. Il montre l'application d'une stratégie alternative dans le cas où les données d'apprentissage sont insuffisantes, mais où de nombreuses données non étiquetées sont à notre disposition : le cotraining multi-classifieurs. Les deux vues indépendantes habituelles du co-training sont remplacées par deux classifieurs basés sur des techniques de classification différentes : icsiboost sur le boosting et LIBLINEAR sur de la régression logistique.

Nous présentons une comparaison de la performance de deux types différents de reconnaisseurs pour le japonais et l'anglais basés sur les grammaires. L'un des systèmes est dérivé à partir de règles d'une grammaire monolingue et l'autre de règles paramétrisées et multilingues. Ce dernier

emploie, les mêmes règles de grammaire pour la création de modèles de langue nécessaires à la reconnaissance des langues typologiquement différentes. Nous avons effectué des expériences sur la reconnaissance dans les applications de dialogue de domaine limitée. Ces expériences montrent que les modèles de langue dérivés des règles multilingues de grammaire (1) traitent aussi bien l'un que l'autre les deux langues examinées, et (2) que leur performance est comparable à celle des reconnaisseurs dérivés de grammaires monolingues. Ceci suggère que le partage de grammaires entre langues typologiquement différentes pourrait être une solution pour rendre plus efficace le développement de systèmes de reconnaissance de la parole linguistiques.

Cette étude s'insère dans le projet VOILADIS (VOIsinage Lexical pour l'Analyse du DIScours), qui a pour objectif d'exploiter des marques de cohésion lexicale pour mettre au jour des phénomènes discursifs. Notre propos est de montrer la pertinence d'une ressource, construite par l'analyse distributionnelle automatique d'un corpus, pour repérer les liens lexicaux dans les textes. Nous désignons par voisins les mots rapprochés par l'analyse distributionnelle sur la base des contextes syntaxiques qu'ils partagent au sein du corpus. Pour évaluer la pertinence de la ressource ainsi créée, nous abordons le problème du repérage des liens lexicaux à travers une application de TAL, la segmentation thématique. Nous discutons l'importance, pour cette tâche, de la ressource lexicale mobilixsée ; puis nous présentons la base de voisins distributionnels que nous utilisons ; enfin, nous montrons qu'elle permet, dans un système de segmentation thématique inspiré de (Hearst, 1997), des performances supérieures à celles obtenues avec une ressource traditionnelle.

La constitution de ressources lexicales est une tâche cruciale pour l'amélioration des performances des systèmes de recherche d'information. Cet article présente une méthode d'extraction d'unités lexicales en chinois contemporain dans un corpus spécialisé non-annoté et non-segmenté. Cette méthode se base sur une construction incrémentale de l'unité lexicale orientée par une mesure d'association. Elle se distingue des travaux précédents par une approche linguistique

non-supervisée assistée par les statistiques. Les résultats de l'extraction, évalués sur un échantillon aléatoire du corpus de travail, sont honorables avec des scores de précision et de rappel respectivement de 52,6 % et 53,7 %.

Les mots sont souvent porteurs de plusieurs sens. Pour traiter l'information correctement, un ordinateur doit être capable de décider quel sens d'un mot est employé à chacune de ses occurrences. Ce problème non parfaitement résolu a généré beaucoup de travaux sur la désambiguïsation du sens des mots (Word Sense Disambiguation) et dans la génération d'espaces sémantiques dont un des buts est de distinguer ces différents sens. Nous nous inspirons ici de deux méthodes existantes de détection automatique des différents usages et/ou sens des mots, pour les appliquer à des espaces sémantiques issus d'une analyse syntaxique effectuée sur un très grand nombre de pages web. Les adaptations et résultats présentés dans cet article se distinguent par le fait d'utiliser non plus une seule représentation mais une combinaison de multiples espaces de forte dimensionnalité. Ces multiples représentations étant en compétition entre elles, elles participent chacune par vote à l'induction des sens lors de la phase de clustering.

Compte tenu de l'essor du Web et du développement des documents multilingues, le besoin de traductions "à la volée" est devenu une évidence. Cet article présente un système qui propose, pour une phrase donnée, non pas une unique traduction, mais une liste de N hypothèses de traductions en faisant appel à plusieurs moteurs de traduction pré-existants. Neufs moteurs de traduction automatique gratuits et disponibles sur leWeb ont été sélectionnés pour soumettre un texte à traduire et réceptionner sa traduction. Les traductions obtenues sont classées selon une métrique reposant sur l'utilisation d'un modèle de langage. Les expériences conduites ont montré que ce méta-moteur de traduction se révèle plus pertinent que l'utilisation d'un seul système de traduction.

La lecture constitue l'une des tâches essentielles dans l'apprentissage d'une langue étrangère.

Toutefois, la découverte d'un texte portant sur un sujet précis et qui soit adapté au niveau de chaque apprenant est consommatrice de temps et pourrait être automatisée. Des expériences montrent que, pour l'anglais, l'utilisation de classifieurs statistiques permet d'estimer automatiquement la difficulté d'un texte. Dans cet article, nous proposons une méthodologie originale comparant, pour le français langue étrangère (FLE), diverses techniques de classification (la régression logistique, le bagging et le boosting) sur deux corpus d'entraînement. Il ressort de cette analyse comparative une légère supériorité de la régression logistique multinomiale.

La normalisation intervient dans de nombreux champs du traitement de l'information. Elle permet d'optimiser les performances des applications, telles que la recherche ou l'extraction d'information, et de rendre plus fiable la constitution de ressources langagières. La normalisation consiste à ramener toutes les variantes d'un même terme ou d'une entité nommée à une forme standard, et permet de limiter l'impact de la variation linguistique. Notre travail porte sur la normalisation des entités nommées, pour laquelle nous avons mis en place un système complexe mêlant plusieurs approches. Nous en présentons ici une des composantes : une méthode endogène de délimitation et de validation de l'entité nommée normée, adaptée à des données multilingues. De plus, nous plaçons l'utilisateur au centre du processus de normalisation, dans l'objectif d'obtenir des données parfaitement fiables et adaptées à ses besoins.

Dans cet article, nous présentons une méthode de transformation de Wikipédia en ressource d'information externe pour détecter et désambiguïser des entités nommées, en milieu ouvert et sans apprentissage spécifique. Nous expliquons comment nous construisons notre système, puis nous utilisons cinq éditions linguistiques de Wikipédia afin d'enrichir son lexique. Pour finir nous réalisons une évaluation et comparons les performances du système avec et sans compléments lexicaux issus des informations inter-linguistiques, sur une tâche d'extraction d'entités nommées appliquée à un corpus d'articles journalistiques.

Nos travaux de recherche s'intéressent à l'application de la théorie de la distance intertextuelle sur la langue arabe en tant qu'outil pour la classification de textes. Cette théorie traite de la classification de textes selon des critères de statistique lexicale, se basant sur la notion de connexion lexicale. Notre objectif est d'intégrer cette théorie en tant qu'outil de classification de textes en langue arabe. Ceci nécessite l'intégration d'une métrique pour la classification de textes au niveau d'une base de corpus lemmatisés étiquetés et identifiés comme étant des références d'époques, de genre, de thèmes littéraires et d'auteurs et ceci afin de permettre la classification de textes anonymes.

Les moyens et les formes stratégiques permettant la génération de descriptions textuelles argumentées d'une même réalité effective sont nombreux. La plupart des définitions proposées de l'argumentation partagent l'idée qu'argumenter c'est fournir les éléments en faveur d'une conclusion donnée. Or dans notre tâche qui consiste à générer des descriptions argumentées pour des accidents de la route, nous ne disposons pas uniquement d'éléments en faveur de la conclusion souhaitée mais aussi d'éléments qui vont à l'encontre de cette dernière et dont la présence est parfois obligatoire pour la compréhension de ces descriptions. Afin de remédier à ce problème, nous proposons des techniques de génération de descriptions argumentées qui présentent au mieux les éléments indésirables à l'aide de stratégies argumentatives.

Cet article présente les résultats d'une évaluation exhaustive des principaux analyseurs syntaxiques probabilistes dit "lexicalisés" initialement conçus pour l'anglais, adaptés pour le français et évalués sur le CORPUS ARBORÉ DU FRANÇAIS (Abeillé et al., 2003) et le MODIFIED FRENCH TREEBANK (Schluter & van Genabith, 2007). Confirmant les résultats de (Crabbé & Candito, 2008), nous montrons que les modèles lexicalisés, à travers les modèles de Charniak (Charniak, 2000), ceux de Collins (Collins, 1999) et le modèle des TIG Stochastiques (Chiang, 2000), présentent des

performances moindres face à un analyseur PCFG à Annotation Latente (Petrov et al., 2006). De plus, nous montrons que le choix d'un jeu d'annotations issus de tel ou tel treebank oriente fortement les résultats d'évaluations tant en constituance qu'en dépendance non typée. Comparés à (Schluter & van Genabith, 2008; Arun & Keller, 2005), tous nos résultats sont state-of-the-art et infirment l'hypothèse d'une difficulté particulière qu'aurait le français en terme d'analyse syntaxique probabiliste et de sources de données.

Cet article aborde deux problèmes d'analyse morpho-sémantique du lexique : (1) attribuer automatiquement une définition à des noms et verbes morphologiquement construits inconnus des dictionnaires mais présents dans les textes ; (2) proposer une analyse combinant règles et analogie, deux techniques généralement contradictoires. Les noms analysés sont apparemment suffixés et composés (HYDROMASSAGE). La plupart d'entre eux, massivement attestés dans les documents (journaux, Internet) sont absents des dictionnaires. Ils sont souvent reliés à des verbes (HYDROMASSER) également néologiques. Le nombre de ces noms et verbes est estimé à 5.400. L'analyse proposée leur attribue une définition par rapport à leur base, et enrichit un lexique de référence pour le TALN au moyen de cette base, si elle est néologique. L'implémentation des contraintes linguistiques qui régissent ces formations est reproductible dans d'autres langues européennes où sont rencontrés les mêmes types de données dont l'analyse reflète le même raisonnement que pour le français.

Cet article propose une méthode pour extraire une analyse en dépendances d'un énoncé à partir de son analyse en constituants avec les grammaires d'interaction. Les grammaires d'interaction sont un formalisme grammatical qui exprime l'interaction entre les mots à l'aide d'un système de polarités. Le mécanisme de composition syntaxique est régi par la saturation des polarités. Les interactions s'effectuent entre les constituants, mais les grammaires étant lexicalisées, ces interactions peuvent se traduire sur les mots. La saturation des polarités lors de l'analyse syntaxique d'un énoncé permet

d'extraire des relations de dépendances entre les mots, chaque dépendance étant réalisée par une saturation. Les structures de dépendances ainsi obtenues peuvent être vues comme un raffinement de l'analyse habituellement effectuée sous forme d'arbre de dépendance. Plus généralement, ce travail apporte un éclairage nouveau sur les liens entre analyse en constituants et analyse en dépendances.

La question de la grammaticalité, et celle duale de l'agrammaticalité, sont des sujets délicats à aborder, dès lors que l'on souhaite intégrer différents degrés, tant de grammaticalité que d'agrammaticalité. En termes d'analyse automatique, les problèmes posés sont de l'ordre de la représentation des connaissances, du traitement, et bien évidement de l'évaluation. Dans cet article, nous nous concentrons sur l'aspect traitement, et nous nous penchons sur la question de l'analyse d'énoncés agrammaticaux. Nous explorons la possibilité de fournir une analyse la plus complète possible pour un énoncé agrammatical, sans l'apport d'information complémentaire telle que par le biais de mal-règles ou autre grammaire d'erreurs. Nous proposons une solution algorithmique qui permet l'analyse automatique d'un énoncé agrammatical, sur la seule base d'une grammaire modèle-théorique de bonne formation. Cet analyseur est prouvé générer une solution optimale, selon un critère numérique maximisé.

Le projet ANNODIS vise la construction d'un corpus de textes annotés au niveau discursif ainsi que le développement d'outils pour l'annotation et l'exploitation de corpus. Les annotations adoptent deux points de vue complémentaires : une perspective ascendante part d'unités de discours minimales pour construire des structures complexes via un jeu de relations de discours ; une perspective descendante aborde le texte dans son entier et se base sur des indices pré-identifiés pour détecter des structures discursives de haut niveau. La construction du corpus est associée à la création de deux interfaces : la première assiste l'annotation manuelle des relations et structures discursives en permettant une visualisation du marquage issu des pré-traitements ; une seconde

sera destinée à l'exploitation des annotations. Nous présentons les modèles et protocoles d'annotation élaborés pour mettre en oeuvre, au travers de l'interface dédiée, la campagne d'annotation.

Cet article présente une série d'évaluations visant à étudier l'apport d'une analyse syntaxique robuste des questions et des documents dans un système de questions-réponses. Ces évaluations ont été effectuées sur le système FIDJI, qui utilise à la fois des informations syntaxiques et des techniques plus "traditionnelles". La sélection des documents, l'extraction de la réponse ainsi que le comportement selon les différents types de questions ont été étudiés.

Le correcteur grammatical Cordial utilise depuis de nombreuses années les co-occurrences pour la désambiguïsation sémantique. Un dictionnaire de co-occurrences ayant été constitué pour les utilisateurs du logiciel de correction et d'aides à la rédaction, la grande richesse de ce dictionnaire a incité à l'utiliser intensivement pour la correction, spécialement des homonymes et paronymes. Les résultats obtenus sont spectaculaires sur ces types d'erreurs mais la prise en compte des co-occurrences a également été utilisée avec profit pour la pure correction orthographique et pour le rattachement des groupes en analyse syntaxique.

La maintenance et l'enrichissement des lexiques morpho-syntaxiques sont souvent des tâches fastidieuses. Dans cet article nous présentons la mise en place d'une procédure de guessing de flexion afin d'aider les linguistes dans leur travail de lexicographes. Le guesser développé ne fait pas qu'évaluer l'étiquette morpho-syntaxique comme c'est généralement le cas. Il propose pour un mot français inconnu, un ou plusieurs candidats-lemmes, ainsi que les paradigmes de flexion associés (formes fléchies et étiquettes morpho-syntaxiques). Dans cet article, nous décrivons le modèle probabiliste utilisé ainsi que les résultats obtenus. La méthode utilisée permet de réduire considérablement le nombre de règles à valider, permettant ainsi un gain de temps important.

Les blogs constituent un support d'observations idéal pour des applications liées à la fouille d'opinion. Toutefois, ils imposent de nouvelles problématiques et de nouveaux défis au regard des méthodes traditionnelles du domaine. De ce fait, nous proposons une méthode automatique pour la détection et la catégorisation des évaluations localement exprimées dans un corpus de blogs multi-domaine. Celle-ci rend compte des spécificités du langage évaluatif décrites dans deux théories linguistiques. L'outil développé au sein de la plateforme UIMA vise d'une part à construire automatiquement une grammaire du langage évaluatif, et d'autre part à utiliser cette grammaire pour la détection et la catégorisation des passages évaluatifs d'un texte. La catégorisation traite en particulier l'aspect axiologique de l'évaluation, sa configuration d'énonciation et sa modalité dans le discours.

Cet article présente la chaîne de traitement linguistique réalisée pour la mise en place d'une plateforme touristique sur Internet. Les premières étapes de cette chaîne sont le repérage et l'annotation des expressions temporelles présentes dans des pages Web. Ces deux tâches sont effectuées à l'aide de patrons linguistiques. Elles soulèvent de nombreux questionnements auxquels nous tentons de répondre, notamment au sujet de la définition des informations à extraire, du format d'annotation et des contraintes. L'étape suivante consiste en l'exploitation des données annotées pour le peuplement d'une ontologie du tourisme. Nous présentons les règles d'acquisition nécessaires pour alimenter la base de connaissance du projet. Enfin, nous exposons une évaluation du système d'annotation. Cette évaluation permet de juger aussi bien le repérage des expressions temporelles que leur annotation.

L'annotation sémantique a pour objectif d'apporter au texte une représentation explicite de son interprétation sémantique. Dans un précédent article, nous avons proposé d'étendre les ontologies par des règles d'annotation sémantique. Ces règles sont utilisées pour l'annotation sémantique d'un

texte au regard d'une ontologie dans le cadre d'une plate-forme d'annotation linguistique automatique. Nous présentons dans cet article une mesure, basée sur la valeur de Shapley, permettant d'identifier les règles qui sont sources de contradiction dans l'annotation sémantique. Par rapport aux classiques mesures de précision et de rappel, l'intérêt de cette mesure est de ne pas nécessiter de corpus manuellement annoté, d'être entièrement automatisable et de permettre l'identification des règles qui posent problème.

Le projet ANR Emotirob aborde la question de la détection des émotions sous un cadre original : concevoir un robot compagnon émotionnel pour enfants fragilisés. Notre approche consiste à combiner détection linguistique et prosodie. Nos expériences montrent qu'un sujet humain peut estimer de manière fiable la valence émotionnelle d'un énoncé à partir de son contenu propositionnel. Nous avons donc développé un premier modèle de détection linguistique qui repose sur le principe de compositionnalité des émotions : les mots simples ont une valence émotionnelle donnée et les prédicats modifient la valence de leurs arguments. Après une description succincte du système logique de compréhension dont les sorties sont utilisées pour le calcul global de l'émotion, cet article présente la construction d'une norme émotionnelle lexicale de référence, ainsi que d'une ontologie de classes émotionnelles de prédicats, pour des enfants de 5 et 7 ans.

Traditionnellement, la morphologie lexicale a été diachronique et a permis de proposer le concept de famille de mots. Ce dernier est repris dans les études en synchronie et repose sur une forte cohérence sémantique entre les mots d'une même famille. Dans cet article, nous proposons une approche en synchronie fondée sur la notion de continuité à la fois phonologique et sémantique. Nous nous intéressons, d'une part, à la morpho-phonologie et, d'autre part, à la dispersion sémantique des mots dans les familles. Une première étude (Gala & Rey, 2008) montrait que les familles de mots obtenues présentaient des espaces sémantiques soit de grande cohésion soit de grande dispersion. Afin de valider ces observations, nous présentons ici une méthode empirique qui

permet de pondérer automatiquement les unités de sens d'un mot et d'une famille. Une expérience menée auprès de 30 locuteurs natifs valide notre approche et ouvre la voie pour une étude approfondie du lexique sur ces bases phonologiques et sémantiques.

Nous présentons des travaux réalisés dans le domaine des systèmes de questions réponses (SQR) utilisant des questions enchaînées. La recherche des documents dans un SQR est perturbée par l'absence des éléments utiles à la recherche dans les questions liées, éléments figurant dans les échanges précédents. Les récentes campagnes d'évaluation montrent que ce problème est sous-estimé, et n'a pas fait l'objet de technique dédiée. Afin d'améliorer la recherche des documents dans un SQR nous utilisons une méthode récente d'organisation des informations liées aux interactions entre questions. Celle-ci se base sur l'exploitation d'une structure de données adaptée à la transmission des informations des questions liées jusqu'au moteur d'interrogation. Le moteur d'interrogation doit alors être adapté afin de tirer partie de cette structure de données.

Cet article présente un moyen de contraindre la production d'expressions référentielles par un système de dialogue en fonction du terrain commun. Cette capacité, fondamentale pour atteindre la compréhension mutuelle, est trop souvent oubliée dans les systèmes de dialogue. Le modèle que nous proposons s'appuie sur une modélisation du processus d'ancrage (grounding process) en proposant un raffinement du statut d'ancrage appliqué à la description des référents. Il décrit quand et comment ce statut doit être révisé en fonction des jugements de compréhension des deux participants ainsi que son influence dans le choix d'une description partagée destinée à la génération d'une expression référentielle.

Dans le présent papier, nous présentons nos travaux sur la gestion du dialogue oral arabe Homme-machine. Ces travaux entrent dans le cadre de la réalisation du serveur vocal interactif SARF (Bahou et al., 2008) offrant des renseignements sur le transport ferroviaire tunisien en langue

arabe standard moderne. Le gestionnaire de dialogue que nous proposons est basé sur une approche structurelle et est composé de deux modèles à savoir, le modèle de tâche et le modèle de dialogue. Le premier modèle permet de i) compléter et vérifier l'incohérence des structures sémantiques représentant les sens utiles des énoncés, ii) générer une requête vers l'application et iii) récupérer le résultat et de formuler une réponse à l'utilisateur en langage naturel. Quant au modèle de dialogue, il assure l'avancement du dialogue avec l'utilisateur et l'identification de ses intentions. L'interaction entre ces deux modèles est assurée grâce à un contexte du dialogue permettant le suivi et la mise à jour de l'historique du dialogue.

Nous présentons un système de correction grammatical ouvert, basé sur des analyses syntaxiques profondes. La spécification grammaticale est une grammaire hors-contexte équipée de structures de traits plates. Après une analyse en forêt partagée où les contraintes d'accord de traits sont relâchées, la détection d'erreur minimise globalement les corrections à effectuer et des phrases alternatives correctes sont automatiquement proposées.

Le langage TimeML a été conçu pour l'annotation des informations temporelles dans les textes, notamment les événements, les expressions de temps et les relations entre les deux. Des consignes d'annotation générales ont été élaborées afin de guider l'annotateur dans cette tâche, mais certains phénomènes linguistiques restent à traiter en détail. Un problème commun dans les tâches de TAL, que ce soit en traduction, en génération ou en compréhension, est celui de l'encodage des constructions à verbe support. Relativement peu d'attention a été portée, jusqu'à maintenant, sur ce problème dans le cadre du langage TimeML. Dans cet article, nous proposons des consignes d'annotation pour les constructions à verbe support.

Dans cet article, nous montrons comment nous avons converti les tables du Lexique-Grammaire en un format TAL, celui du lexique Lefff, permettant ainsi son intégration dans l'analyseur syntaxique

FRMG. Nous présentons les fondements linguistiques de ce processus de conversion et le lexique obtenu. Nous validons le lexique obtenu en évaluant l'analyseur syntaxique FRMG sur le corpus de référence de la campagne EASy selon qu'il utilise les entrées verbales du Lefff ou celles des tables des verbes du Lexique-Grammaire ainsi converties.

Les ressources lexicales sont essentielles pour obtenir des systèmes de traitement des langues performants. Ces ressources peuvent être soit construites à la main, soit acquises automatiquement à partir de gros corpus. Dans cet article, nous montrons la complémentarité de ces deux approches. Pour ce faire, nous utilisons l'exemple de la sous-catégorisation verbale en comparant un lexique acquis par des méthodes automatiques (LexSchem) avec un lexique construit manuellement (Le Lexique-Grammaire). Nous montrons que les informations acquises par ces deux méthodes sont bien distinctes et qu'elles peuvent s'enrichir mutuellement.

Cordial est un analyseur syntaxique et sémantique développé par la société Synapse Développement. Largement utilisé par les laboratoires de TALN depuis plus de dix ans, cet analyseur participe à la campagne Passage ("Produire des Annotations Syntaxiques à Grande Échelle"). Comment fonctionne cet analyseur ? Quels résultats a-t-il obtenu lors de la première phase d'évaluation de cette campagne ? Au-delà de ces questions, cet article montre en quoi les contraintes industrielles façonnent les outils d'analyse automatique du langage naturel.

La nécessité d'une interaction systématique entre modèles, traitements et corpus impose la disponibilité d'annotations de référence auxquelles modèles et traitements pourront être confrontés. Or l'établissement de telles annotations requiert un cadre formel permettant la représentation d'objets linguistiques variés, et des applications permettant à l'annotateur de localiser sur corpus et de caractériser les occurrences des phénomènes observés. Si différents outils d'annotation ont vu le jour, ils demeurent souvent fortement liés à un modèle théorique et à des objets linguistiques

particuliers, et ne permettent que marginalement d'explorer certaines structures plus récemment appréhendées expérimentalement, notamment à granularité élevée et en matière d'analyse du discours. La plate-forme Glozz répond à ces différentes contraintes et propose un environnement d'exploration de corpus et d'annotation fortement configurable et non limité a priori au contexte discursif dans lequel elle a initialement vu le jour.

Le traitement automatique des langues exige un recensement lexical aussi rigoureux que possible. Dans ce but, nous avons développé un dictionnaire morphologique du français, conçu comme le point de départ d'un système modulaire (Morfetik) incluant un moteur de flexion, des interfaces de consultation et d'interrogation et des outils d'exploitation. Nous présentons dans cet article, après une brève description du dictionnaire de base (lexique des mots simples), quelques-uns des outils informatiques liés à cette ressource : un moteur de recherche des lemmes et des formes fléchies ; un moteur de flexion XML et MySQL ; des outils NLP permettant d'exploiter le dictionnaire ainsi généré ; nous présentons notamment un analyseur linguistique développé dans notre laboratoire. Nous comparons dans une dernière partie Morfetik avec d'autres ressources analogues du français : Morphalou, Lexique3 et le DELAF.

Dans cet article nous nous intéressons au problème de la détection de réutilisation de texte. Plus particulièrement, étant donné un document original et un ensemble de documents candidats ? thématiquement similaires au premier ? nous cherchons à classer ceux qui sont dérivés du document original et ceux qui ne le sont pas. Nous abordons le problème selon deux approches : dans la première, nous nous intéressons aux similarités discursives entre les documents, dans la seconde au recouvrement de n-grams hapax. Nous présentons le résultat d'expérimentations menées sur un corpus de presse francophone construit dans le cadre du projet ANR PIITHIE.

Notre travail se situe dans le cadre des systèmes de réponse a une question et à pour but de fournir

une réponse en langue naturelle aux questions posées en langue naturelle. Cet article présente une expérience permettant d'analyser les réponses de locuteurs du français à des questions que nous leur posons. L'expérience se déroule à l'écrit comme à l'oral et propose à des locuteurs français des questions relevant de différents types sémantiques et syntaxiques. Nous mettons en valeur une large variabilité dans les formes de réponses possibles en langue française. D'autre part nous établissons un certain nombre de liens entre formulation de question et formulation de réponse. Nous proposons d'autre part une comparaison des réponses selon la modalité oral / écrit. Ces résultats peuvent être intégrés à des systèmes existants pour produire une réponse en langue naturelle de façon dynamique.

La phonétisation est une étape essentielle pour le traitement de l'oral. Dans cet article, nous décrivons un système automatique de phonétisation de mots isolés qui est simple, portable et performant. Il repose sur une approche par apprentissage ; le système est donc construit à partir d'exemples de mots et de leur représentation phonétique. Nous utilisons pour cela une technique d'inférence de règles de réécriture initialement développée pour la translittération et la traduction. Pour évaluer les performances de notre approche, nous avons utilisé plusieurs jeux de données couvrant différentes langues et divers alphabets phonétiques, tirés du challenge Pascal Pronalsyl. Les très bons résultats obtenus égalent ou dépassent ceux des meilleurs systèmes de l'état de l'art.

Les systèmes de traduction statistiques intègrent différents types de modèles dont les prédictions sont combinées, lors du décodage, afin de produire les meilleures traductions possibles. Traduire correctement des mots polysémiques, comme, par exemple, le mot avocat du français vers l'anglais (lawyer ou avocado), requiert l'utilisation de modèles supplémentaires, dont l'estimation et l'intégration s'avèrent complexes. Une alternative consiste à tirer parti de l'observation selon laquelle les ambiguïtés liées à la polysémie ne sont pas les mêmes selon les langues source considérées. Si l'on dispose, par exemple, d'une traduction vers l'espagnol dans laquelle avocat a été traduit par

aguacate, alors la traduction de ce mot vers l'anglais n'est plus ambiguë. Ainsi, la connaissance d'une traduction français!espagnol permet de renforcer la sélection de la traduction avocado pour le système français!anglais. Dans cet article, nous proposons d'utiliser des documents en plusieurs langues pour renforcer les choix lexicaux effectués par un système de traduction automatique. En particulier, nous montrons une amélioration des performances sur plusieurs métriques lorsque les traductions auxiliaires utilisées sont obtenues manuellement.

L'objectif du travail présenté ici est la modélisation de la détection et la correction des erreurs orthographiques et dactylographiques, plus particulièrement dans le contexte des handicaps langagiers. Le travail est fondé sur une analyse fine des erreurs d'écriture commises. La première partie de cet article est consacrée à une description précise de la faute. Dans la seconde partie, nous analysons l'erreur (1) en déterminant la nature de la faute (typographique, orthographique, ou grammaticale) et (2) en explicitant sa conséquence sur le niveau de perturbation linguistique (phonologique, orthographique, morphologique ou syntaxique). Il résulte de ce travail un modèle général des erreurs (une grille) que nous présenterons, ainsi que les résultats statistiques correspondants. Enfin, nous montrerons sur des exemples, l'utilité de l'apport de cette grille, en soumettant ces types de fautes à quelques correcteurs. Nous envisageons également les implications informatiques de ce travail.

Le marché d'offres d'emploi et des candidatures sur Internet connaît une croissance exponentielle. Ceci implique des volumes d'information (majoritairement sous la forme de texte libre) qu'il n'est plus possible de traiter manuellement. Une analyse et catégorisation assistées nous semble pertinente en réponse à cette problématique. Nous proposons E-Gen, système qui a pour but l'analyse et catégorisation assistés d'offres d'emploi et des réponses des candidats. Dans cet article nous présentons plusieurs stratégies, reposant sur les modèles vectoriel et probabiliste, afin de résoudre la problématique du profilage des candidatures en fonction d'une offre précise. Nous

avons évalué une palette de mesures de similarité afin d'effectuer un classement pertinent des candidatures au moyen des courbes ROC. L'utilisation d'une forme de relevance feed-back a permis de surpasser nos résultats sur ce problème difficile et sujet à une grande subjectivité.

Le calcul de la similarité sémantique entre les termes repose sur l'existence et l'utilisation de ressources sémantiques. Cependant de telles ressources, qui proposent des équivalences entre entités, souvent des relations de synonymie, doivent elles-mêmes être d'abord analysées afin de définir des zones de fiabilité où la similarité sémantique est plus forte. Nous proposons une méthode d'acquisition de synonymes élémentaires grâce à l'exploitation des terminologies structurées au travers l'analyse de la structure syntaxique des termes complexes et de leur compositionnalité. Les synonymes acquis sont ensuite profilés grâce aux indicateurs endogènes inférés automatiquement à partir de ces mêmes terminologies (d'autres types de relations, inclusions lexicales, productivité, forme des composantes connexes). Dans le domaine biomédical, il existe de nombreuses terminologies structurées qui peuvent être exploitées pour la constitution de ressources sémantiques. Le travail présenté ici exploite une de ces terminologies, Gene Ontology.

Dans le cadre dune approche déterministe et incrémentale danalyse syntaxique par classification de textes en langue arabe, nous avons prévu de prendre en considération un ensemble varié dattributs discriminants afin de mieux assister la procédure de classification dans ses prises de décisions à travers les différentes étapes danalyse. Ainsi, en plus des attributs morpho-syntaxiques du mot en cours danalyse et des informations contextuelles des mots lavoisinant, nous avons ajouté des informations compositionnelles extraites du fragment de larbre syntaxique déjà construit lors de létape précédente de lanalyse en cours. Ce papier présente notre approche danalyse syntaxique par classification et vise lexposition dune justification expérimentale de lapport de chaque type dattributs discriminants et spécialement ceux compositionnels dans ladite analyse syntaxique.

Cet article présente Recto /Verso, un système de traitement automatique du langage dédié à l'application des rectifications orthographiques de 1990. Ce système a été développé dans le cadre de la campagne de sensibilisation réalisée en mars dernier par le Service et le Conseil de la langue française et de la politique linguistique de la Communauté française de Belgique. Nous commençons par rappeler les motivations et le contenu de la réforme proposée, et faisons le point sur les principes didactiques retenus dans le cadre de la campagne. La plus grande partie de l'article est ensuite consacrée à l'implémentation du système. Nous terminons enfin par une première analyse de l'impact de la campagne sur les utilisateurs.

L'accès aux documents multimédia, dans une archive audiovisuelle, dépend en grande partie de la quantité et de la qualité des méta-données attachées aux documents, notamment la description de leur contenu. Cependant, l'annotation manuelle des collections est astreignante pour le personnel. De nombreuses archives évoluent vers des méthodes d'annotation (semi-)automatiques pour la création et/ou l'amélioration des méta-données. Le project CATCH-CHOICE, fondé par NWO, s'est penché sur l'extraction de mots clés à partir de resources textuelles liées aux programmes TV destinés à être archivés (péritextes), en collaboration avec les archives audiovisuelles néerlandaises, Sound and Vision. Cet article se penche sur la question de l'adéquation des transcriptions de Reconnaissance Automatique de la Parole développés dans le projet CATCH-CHoral pour la génération automatique de mots-clés : les mots-clés extraits de ces ressources sont évalués par rapport à des annotations manuelles et par rapport à des mots-clés générés à partir de péritextes décrivant les programmes télévisuels.

Le résumé automatique de texte est une problématique difficile, fortement dépendante de la langue et qui peut nécessiter un ensemble de données d'apprentissage conséquent. L'approche par extraction peut aider à surmonter ces difficultés. (Mihalcea, 2004) a démontré l'intérêt des approches à base de graphes pour l'extraction de segments de texte importants. Dans cette étude,

nous décrivons une approche indépendante de la langue pour la problématique du résumé automatique multi-documents. L'originalité de notre méthode repose sur l'utilisation d'une mesure de similarité permettant le rapprochement de segments morphologiquement proches. De plus, c'est à notre connaissance la première fois que l'évaluation d'une approche de résumé automatique multi-document est conduite sur des textes en français.

Cette étude porte sur l'analyse de conversations entre des clients et des téléconseillers d'EDF. Elle propose une chaîne de traitements permettant d'automatiser la détection des sujets abordés dans chaque conversation. L'aspect multi-thématique des conversations nous incite à trouver une unité de documents entre le simple tour de parole et la conversation entière. Cette démarche enchaîne une étape de segmentation de la conversation en thèmes homogènes basée sur la notion de cohésion lexicale, puis une étape de text-mining comportant une analyse linguistique enrichie d'un vocabulaire métier spécifique à EDF, et enfin une classification non supervisée des segments obtenus. Plusieurs algorithmes de segmentation ont été évalués sur un corpus de test, segmenté et annoté manuellement : le plus « proche » de la segmentation de référence est C99. Cette démarche, appliquée à la fois sur un corpus de conversations transcrites à la main, et sur les mêmes conversations décodées par un moteur de reconnaissance vocale, aboutit quasiment à l'obtention des 20 mêmes classes thématiques.

Dans cet article, nous traitons du problème de la modélisation statistique du langage pour les langues peu dotées et sans segmentation entre les mots. Tandis que le manque de données textuelles a un impact sur la performance des modèles, les erreurs introduites par la segmentation automatique peuvent rendre ces données encore moins exploitables. Pour exploiter au mieux les données textuelles, nous proposons une méthode qui effectue des segmentations multiples sur le corpus d'apprentissage au lieu d'une segmentation unique. Cette méthode basée sur les automates d'état finis permet de retrouver les n-grammes non trouvés par la segmentation unique et de

générer des nouveaux n-grammes pour l'apprentissage de modèle du langage. L'application de cette approche pour l'apprentissage des modèles de langage pour les systèmes de reconnaissance automatique de la parole en langue khmère et vietnamienne s'est montrée plus performante que la méthode par segmentation unique, à base de règles.

L'obtention d'informations lexicales fiables est un enjeu primordial en TALN, mais cette collecte peut s'avérer difficile. L'approche présentée ici vise à pallier les écueils de cette difficulté en faisant participer un grand nombre de personnes à un projet contributif via des jeux accessibles sur le web. Ainsi, les joueurs vont construire le réseau lexical, en fournissant de plusieurs manières possibles des associations de termes à partir d'un terme cible et d'une consigne correspondant à une relation typée. Le réseau lexical ainsi produit est de grande taille et comporte une trentaine de types de relations. A partir de cette ressource, nous abordons la question de la détermination des différents sens et usages d'un terme. Ceci est réalisé en analysant les relations entre ce terme et ses voisins immédiats dans le réseau et en calculant des cliques ou des quasi-cliques. Ceci nous amène naturellement à introduire la notion de similarité entre cliques, que nous interprétons comme une mesure de similarité entre ces différents sens et usages. Nous pouvons ainsi construire pour un terme son arbre des usages, qui est une structure de données exploitable en désambiguïsation de sens. Nous présentons quelques résultats obtenus en soulignant leur caractère évolutif.

Cet article présente un travail de modélisation et de détection des phénomènes de disfluence. Une des spécificité de ce travail est le cadre dans lequel il se situe: le contrôle de la navigation aérienne. Nous montrons ce que ce cadre particulier implique certains choix concernant la modélisation et l'implémentation. Ainsi, nous constatons que la modélisation fondée sur la syntaxe, souvent utilisée dans le traitement des langues naturelles, n'est pas la plus appropriée ici. Nous expliquons la façon dont l'implémentation a été réalisée. Dans une dernière partie, nous présentons la validation de ce dispositif, effectuée sur 400 énoncés.

Nous présentons ici différents algorithmes d'analyse pour grammaires à concaténation d'intervalles (Range Concatenation Grammar, RCG), dont un nouvel algorithme de type Earley, dans le paradigme de l'analyse déductive. Notre travail est motivé par l'intérêt porté récemment à ce type de grammaire, et comble un manque dans la littérature existante.

Cet article décrit un modèle de langage dédié au dialogue homme-machine, son implémentation en CLIPS, ainsi qu'une évaluation comparative. Notre problématique n'est ni d'analyser des grands corpus, ni de proposer une grammaire à grande couverture. Notre objectif est de produire des représentations sémantiques utilisables par un module de dialogue à partir d'énoncés oraux courts, le plus souvent agrammaticaux. Une démarche pragmatique nous a conduit à fonder l'analyse sur des principes simples mais efficaces dans le cadre que nous nous sommes fixé. L'algorithme retenu s'inspire de l'analyse tabulaire. L'évaluation que nous présentons repose sur le corpus MEDIA qui a fait l'objet d'une annotation sémantique manuelle de référence pour une campagne d'évaluation d'analyseurs sémantiques pour le dialogue. Les résultats que nous obtenons place notre analyseur dans le trio de tête des systèmes évalués lors de la campagne de juin 2005, et nous confortent dans nos choix d'algorithme et de représentation des connaissances.

Nous présentons une approche exploratoire basée sur des notions thermodynamiques de la Physique statistique pour la compression de phrases. Nous décrivons le modèle magnétique des verres de spins, adapté à notre conception de la problématique. Des simulations Métropolis Monte-Carlo permettent d'introduire des fluctuations thermiques pour piloter la compression. Des comparaisons intéressantes de notre méthode ont été réalisées sur un corpus en français.

Les schémas de pondération utilisés habituellement en catégorisation de textes, et plus généralement en recherche d'information (RI), ne sont pas adaptés à l'utilisation de données liées à

des textes issus d'un processus de reconnaissance de l'écriture. En particulier, les candidats-mot à la reconnaissance ne pourraient être exploités sans introduire de fausses occurrences de termes dans le document. Dans cet article nous présentons un nouveau schéma de pondération permettant d'exploiter les listes de candidats-mot. Il permet d'estimer le pouvoir discriminant d'un terme en fonction de la probabilité a posteriori d'un candidat-mot dans une liste de candidats. Les résultats montrent que le taux de classification de documents fortement dégradés peut être amélioré en utilisant le schéma proposé.

Contrairement à la plupart des systèmes de traitement du langage, qui s'appliquent à des langues écrites et standardisées, nous présentons ici un système de traduction automatique qui prend en compte les spécificités des dialectes. En général, les dialectes se caractérisent par une variation continue et un manque de données textuelles en qualité et quantité suffisantes. En même temps, du moins en Europe, les dialectologues ont étudié en détail les caractéristiques linguistiques des dialectes. Nous soutenons que des données provenant d'atlas dialectologiques peuvent être utilisées pour paramétrer un système de traduction automatique. Nous illustrons cette idée avec le prototype d'un système de traduction basé sur des règles, qui traduit de l'allemand standard vers les différents dialectes de Suisse allemande. Quelques exemples linguistiquement motivés serviront à exposer l'architecture de ce système.

Le modèle PLSI (« Probabilistic Latent semantic Indexing ») offre une approche de l'indexation de documents fondée sur des modèles probabilistes de catégories sémantiques latentes et a conduit à des applications dans différents domaines. Toutefois, ce modèle rend impossible le traitement de documents inconnus au moment de l'apprentissage, problème particulièrement sensible pour la représentation des requêtes dans le cadre de la recherche d'information. Une méthode, dite de « folding-in », permet dans une certaine mesure de contourner ce problème, mais présente des faiblesses. Cet article introduit nouvelle une mesure de similarité document-requête pour PLSI,

fondée sur lesmodèles de langue, où le problème du « folding-in » ne se pose pas. Nous comparons cette nouvelle similarité aux noyaux de Fisher, l'état de l'art en la matière. Nous présentons aussi une évaluation de PLSI sur un corpus de recherche d'information de près de 7500 documents et de plus d'un million d'occurrences de termes provenant de la collection TREC?AP, une taille considérable dans le cadre de PLSI.

La segmentation thématique est un domaine de l'analyse discursive ayant donné lieu à de nombreux travaux s'appuyant sur la notion de cohésion lexicale. La plupart d'entre eux n'exploitent que la simple récurrence lexicale mais quelques uns ont néanmoins exploré l'usage de connaissances rendant compte de cette cohésion lexicale. Celles-ci prennent généralement la forme de réseaux lexicaux, soit construits automatiquement à partir de corpus, soit issus de dictionnaires élaborés manuellement. Dans cet article, nous examinons dans quelle mesure une ressource d'une nature un peu différente peut être utilisée pour caractériser la cohésion lexicale des textes. Il s'agit en l'occurrence de sens de mots induits automatiquement à partir de corpus, à l'instar de ceux produits par la tâche «Word Sense Induction and Discrimination » de l'évaluation SemEval 2007. Ce type de ressources apporte une structuration des réseaux lexicaux au niveau sémantique dont nous évaluons l'apport pour la segmentation thématique.

Dans cette démonstration, nous présentons le prototype d'un environnement open-source pour l'édition de corpus de dépendances. Cet environnement, nommé ACOLAD (Annotation de COrpus Linguistique pour l'Analyse de dépendances), propose des services manuels de segmentation et d'annotation multi-niveaux (segmentation en mots et en syntagmes minimaux (chunks), annotation morpho-syntaxique des mots, annotation syntaxique des chunks et annotation syntaxique des dépendances entre mots ou entre chunks).

Dans cet article, nous présentons une application sur le web pour l'acquisition de paraphrases

phrastiques et sous-phrastiques sous forme de jeu. L'application permet l'acquisition à la fois de paraphrases et de jugements humains multiples sur ces paraphrases, ce qui constitue des données particulièrement utiles pour les applications du TAL basées sur les phénomènes paraphrastiques.

Nous présentons anymalign, un aligneur sous-phrastique grand public. Ses résultats ont une qualité qui rivalise avec le meilleur outil du domaine, GIZA++. Il est rapide et simple d'utilisation, et permet de produire dictionnaires et autres tables de traduction en une seule commande. À notre connaissance, c'est le seul outil au monde permettant d'aligner un nombre quelconque de langues simultanément. Il s'agit donc du premier aligneur sous-phrastique réellement multi-lingue.

La construction d'ontologie à partir de textes fait l'objet d'études depuis plusieurs années dans le domaine de l'ingénierie des ontologies. Un cadre méthodologique en quatre étapes (constitution d'un corpus de documents, analyse linguistique du corpus, conceptualisation, opérationnalisation de l'ontologie) est commun à la plupart des méthodes de construction d'ontologies à partir de textes. S'il existe plusieurs plateformes de traitement automatique de la langue (TAL) permettant d'analyser automatiquement les corpus et de les annoter tant du point de vue syntaxique que statistique, il n'existe actuellement aucune procédure généralement acceptée, ni a fortiori aucun ensemble cohérent d'outils supports, permettant de concevoir de façon progressive, explicite et traçable une ontologie de domaine à partir d'un ensemble de ressources informationnelles relevant de ce domaine. Le but de ce court article est de présenter les propositions développées, au sein du projet ANR DaFOE 4app, pour favoriser l'émergence d'un tel ensemble d'outils.

L'analyse qualitative des données demande au sociologue un important travail de sélection et d'interprétation des documents. Afin de faciliter ce travail, cette communauté c'est dotée d'outils informatique mais leur fonctionnalités sont encore limitées. Le projet ASSIST est une étude exploratoire pour préciser les modules de traitement automatique des langues (TAL) permettant

d'assister le sociologue dans son travail d'analyse. Nous présentons le moteur de recherche réalisé et nous justifions le choix des composants de TAL intégrés au prototype.

CETLEF.fr ? une application Web dynamique ? propose des exercices de déclinaison tchèque avec un diagnostic automatique des erreurs. Le diagnostic a nécessité l'élaboration d'un modèle formel spécifique de la déclinaison contenant un classement des types paradigmatiques et des règles pour la réalisation des alternances morphématiques. Ce modèle est employé pour l'annotation des formes requises, nécessaire pour le diagnostic, mais également pour une présentation didactique sur la plateforme apprenant. Le diagnostic est effectué par comparaison d'une production erronée avec des formes hypothétiques générées à partir du radical de la forme requise et des différentes désinences casuelles. S'il existe une correspondance, l'erreur est interprétée d'après les différences dans les traits morphologiques de la forme requise et de la forme hypothétique. La majorité des erreurs commises peut être interprétée à l'aide de cette technique.

CIFLI-SurviTra ("Survival Translation" assistant) est une plate-forme destinée à favoriser l'ingénierie et la mise au point de composants UNL de TA, à partir d'une mémoire de traduction formée de livres de phrases multi-lingues avec variables lexicales. SurviTra est aussi un phrasebook digital multi-lingue, assistant linguistique pour voyageurs monolingues (français, hindi, tamoul, anglais) en situation de "survie linguistique". Le corpus d'un domaine-pilote ("Restaurant") a été structuré et construit : sous-domaines de phrases alignées et classes lexicales de locutions quadrilingues, graphes UNL, dictionnaires UW++/français et UW++/hindi par domaines. L'approche, générique, est applicable à d'autres langues. Le prototype d'assistant linguistique (application Web, à interface textuelle) peut évoluer vers une application UNL embarquée sur SmartPhone, avec Traitement de Parole et multimodalité.

Nous présentons ici PolArt, un outil multi-lingue pour l'analyse de sentiments qui aborde la

composition des sentiments en appliquant des transducteurs en cascade. La compositionnalité est assurée au moyen de polarités préalables extraites d'un lexique et des règles de composition appliquées de manière incrémentielle.

Nous proposons une plateforme d?annotation sémantique, appelée « EXCOM ». Basée sur la méthode de l? « Exploration Contextuelle », elle permet, à travers une diversité de langues, de procéder à des annotations automatiques de segments textuels par l'analyse des formes de surface dans leur contexte. Les textes sont traités selon des « points de vue » discursifs dont les valeurs sont organisées dans une « carte sémantique ». L?annotation se base sur un ensemble de règles linguistiques, écrites par un analyste, qui permettent d?identifier les représentations textuelles sous-jacentes aux différentes catégories de la carte. Le système offre, à travers deux types d?interfaces (développeur ou utilisateur), une chaîne de traitements automatiques de textes qui comprend la segmentation, l?annotation et d?autres fonctionnalités de post-traitement. Les documents annotés peuvent être utilisés, par exemple, pour des systèmes de recherche d?information, de veille, de classification ou de résumé automatique.

Nous présentons la dernière version du logiciel SAGACE, analyseur de corpus pour langues faiblement flexionnelles (par exemple japonais ou chinois). Ce logiciel est distribué avec un lexique où les catégories sont exprimées à l'aide de systèmes de traits.

L'objectif de la démonstration est d'une part de faire un retour d'expérience sur la solution logicielle Apache UIMA comme infrastructure de développement d'applications distribuées de TAL, et d'autre part de présenter les développements réalisés par l'équipe TALN du LINA pour permettre à la communauté de s'approprier ce « framework ».

Le chunking consiste à segmenter un texte en chunks, segments sous-phrastiques qu'Abney a

défini approximativement comme des groupes accentuels. Traditionnellement, le chunking utilise des ressources monolingues, le plus souvent exhaustives, quelquefois partielles : des mots grammaticaux et des ponctuations, qui marquent souvent des débuts et fins de chunk. Mais cette méthode, si l'on veut l'étendre à de nombreuses langues, nécessite de multiplier les ressources monolingues. Nous présentons une nouvelle méthode : le chunking endogène, qui n'utilise aucune ressource hormis le texte analysé lui-même. Cette méthode prolonge les travaux de Zipf : la minimisation de l'effort de communication conduit les locuteurs à raccourcir les mots fréquents. On peut alors caractériser un chunk comme étant la période des fonctions périodiques correllées longueur et effectif des mots sur l'axe syntagmatique. Cette méthode originale présente l'avantage de s'appliquer à un grand nombre de langues d'écriture alphabétique, avec le même algorithme, sans aucune ressource.

L'article propose un modèle linguistique et informatique permettant de faire émerger la structure morphologique dérivationnelle du lexique à partir des régularités sémantiques et formelles des mots qu'il contient. Ce modèle est radicalement lexématique. La structure morphologique est constituée par les relations que chaque mot entretient avec les autres unités du lexique et notamment avec les mots de sa famille morphologique et de sa série dérivationnelle. Ces relations forment des paradigmes analogiques. La modélisation a été testée sur le lexique du français en utilisant le dictionnaire informatisé TLFi.

Nous proposons un algorithme d'analyse pour les grammaires d'interaction qui utilise le cadre formel de l'analyse déductive. Cette approche donne un point de vue nouveau sur ce problème puisque les méthodes précédentes réduisaient ce dernier à la réécriture de graphes et utilisaient des techniques de résolution de contraintes. D'autre part, cette présentation permet de décrire le processus de manière standard et d'exhiber les sources d'indéterminisme qui rendent ce problème difficile.

Cet article décrit un modèle d'analyse syntaxique de l'oral spontané axé sur la reconnaissance de cadres valenciels verbaux. Le modèle d'analyse se décompose en deux étapes : une étape générique, basée sur des ressources génériques du français et une étape de ré-ordonnancement des solutions de l'analyseur réalisé par un modèle spécifique à une application. Le modèle est évalué sur le corpus MEDIA.

Cet article présente une technique d'analyse syntaxique statistique à la fois en constituants et en dépendances. L'analyse procède en ajoutant des étiquettes fonctionnelles aux sorties d'un analyseur en constituants, entraîné sur le French Treebank, pour permettre l'extraction de dépendances typées. D'une part, nous spécifions d'un point de vue formel et linguistique les structures de dépendances à produire, ainsi que la procédure de conversion du corpus en constituants (le French Treebank) vers un corpus cible annoté en dépendances, et partiellement validé. D'autre part, nous décrivons l'approche algorithmique qui permet de réaliser automatiquement le typage des dépendances. En particulier, nous nous focalisons sur les méthodes d'apprentissage discriminantes d'étiquetage en fonctions grammaticales.

L'objectif de cet article est d'évaluer dans quelle mesure les "fonctions syntaxiques" qui figurent dans une partie du corpus arboré de Paris 7 sont apprenables à partir d'exemples. La technique d'apprentissage automatique employée pour cela fait appel aux "Champs Aléatoires Conditionnels" (Conditional Random Fields ou CRF), dans une variante adaptée à l'annotation d'arbres. Les expériences menées sont décrites en détail et analysées. Moyennant un bon paramétrage, elles atteignent une F1-mesure de plus de 80%.

Les principaux travaux en extraction de lexiques bilingues à partir de corpus comparables reposent sur l'hypothèse implicite que ces corpus sont équilibrés. Cependant, les différentes méthodes

computationnelles associées sont relativement insensibles à la taille de chaque partie du corpus. Dans ce contexte, nous étudions l'influence que peut avoir un corpus comparable déséquilibré sur la qualité des terminologies bilingues extraites à travers différentes expériences. Nos résultats montrent que sous certaines conditions l'utilisation d'un corpus comparable déséquilibré peut engendrer un gain significatif dans la qualité des lexiques extraits.

On utilise souvent des ressources lexicales externes pour améliorer les performances des systèmes d'étiquetage d'entités nommées. Les contenus de ces ressources lexicales peuvent être variés : liste de noms propres, de lieux, de marques. On note cependant que la disponibilité de corpus encyclopédiques exhaustifs et ouverts de grande taille tels que Worldnet ou Wikipedia, a fait émerger de nombreuses propositions spécifiques d'exploitation de ces contenus par des systèmes d'étiquetage. Un problème demeure néanmoins ouvert avec ces ressources : celui de l'adaptation de leur taxonomie interne, complexe et composée de dizaines de milliers catégories, aux exigences particulières de l'étiquetage des entités nommées. Pour ces dernières, au plus de quelques centaines de classes sémantiques sont requises. Dans cet article nous explorons cette difficulté et proposons un système complet de transformation d'un arbre taxonomique encyclopédique en une système à classe sémantiques adapté à l'étiquetage d'entités nommées.

Plusieurs travaux ont récemment étudié l'apport de l'apprentissage analogique dans des applications du traitement automatique des langues comme la traduction automatique, ou la recherche d'information. Il est souvent admis que les relations analogiques de forme entre les mots capturent des informations de nature morphologique. Le but de cette étude est de présenter une analyse des points de rencontre entre l'analyse morphologique et les analogies de forme. C'est à notre connaissance la première étude de ce type portant sur des corpus de grande taille et sur plusieurs langues. Bien que notre étude ne soit pas dédiée à une tâche particulière du traitement des langues, nous montrons cependant que le principe d'analogie permet de segmenter des mots

en morphèmes avec une bonne précision.

Cet article présente nos premiers travaux en vue de la construction d'un système de traduction probabiliste pour le couple de langue vietnamien-français. La langue vietnamienne étant considérée comme une langue peu dotée, une des difficultés réside dans la constitution des corpus parallèles, indispensable à l'apprentissage des modèles. Nous nous concentrons sur la constitution d'un grand corpus parallèle vietnamien-français. La méthode d'identification automatique des paires de documents parallèles fondée sur la date de publication, les mots spéciaux et les scores d'alignements des phrases est appliquée. Cet article présente également la construction d'un premier système de traduction automatique probabiliste vietnamien-français et français-vietnamien à partir de ce corpus et discute l'opportunité d'utiliser des unités lexicales ou sous-lexicales pour le vietnamien (syllabes, mots, ou leurs combinaisons). Les performances du système sont encourageantes et se comparent avantageusement à celles du système de Google.

L'extraction de lexiques bilingues à partir de corpus comparables affiche de bonnes performances pour des corpus volumineux mais chute fortement pour des corpus d'une taille plus modeste. Pour pallier cette faiblesse, nous proposons une nouvelle contribution au processus d'alignement lexical à partir de corpus comparables spécialisés qui vise à renforcer la significativité des contextes lexicaux en s'appuyant sur le vocabulaire spécialisé du domaine étudié. Les expériences que nous avons réalisées en ce sens montrent qu'une meilleure prise en compte du vocabulaire spécialisé permet d'améliorer la qualité des lexiques extraits.

Malgré les nombreuses études visant à améliorer la traduction automatique, la traduction assistée par ordinateur reste la solution préférée des traducteurs lorsqu'une sortie de qualité est recherchée. Dans cet article, nous présentons nos travaux menés dans le but d'améliorer le concordancier bilingue TransSearch. Ce service, accessible sur le Web, repose principalement sur un alignement

au niveau des phrases. Dans cette étude, nous discutons et évaluons l'intégration d'un alignement statistique au niveau des mots. Nous présentons deux nouvelles problématiques essentielles au succès de notre nouveau prototype : la détection des traductions erronées et le regroupement des variantes de traduction similaires.

L'article présente une étude portant sur des constituants détachés à valeur axiologique. Dans un premier temps, une analyse linguistique sur corpus met en évidence un ensemble de patrons caractéristiques du phénomène. Ensuite, une expérimentation informatique est proposée sur un corpus de plus grande taille afin de permettre l'observation des patrons en vue d'un retour sur le modèle linguistique. Ce travail s'inscrit dans un projet mené à l'interface de la linguistique et du TAL, qui se donne pour but d'enrichir, d'adapter au français et de formaliser le modèle général Appraisal de l'évaluation dans la langue.

Notre société génère une masse d'information toujours croissante, que ce soit en médecine, en météorologie, etc. La méthode la plus employée pour analyser ces données est de les résumer sous forme graphique. Cependant, il a été démontré qu'un résumé textuel est aussi un mode de présentation efficace. L'objectif du prototype BT-45, développé dans le cadre du projet Babytalk, est de générer des résumés de 45 minutes de signaux physiologiques continus et d'événements temporels discrets en unité néonatale de soins intensifs (NICU). L'article présente l'aspect génération de texte de ce prototype. Une expérimentation clinique a montré que les résumés humains améliorent la prise de décision par rapport à l'approche graphique, tandis que les textes de BT-45 donnent des résultats similaires à l'approche graphique. Une analyse a identifié certaines des limitations de BT-45 mais en dépit de celles-ci, notre travail montre qu'il est possible de produire automatiquement des résumés textuels efficaces de données complexes.

Dans cet article, nous nous intéressons aux adjectifs dits relationnels et à leur statut en traitement

automatique des langues naturelles (TALN). Nous montrons qu'ils constituent une « sous-classe » d'adjectifs rarement explicitée et donc rarement représentée dans les lexiques sur lesquels reposent les applications du TALN, alors qu'ils jouent un rôle important dans de nombreuses applications. Leur formation morphologique est source d'importantes divergences entre différentes langues, et c'est pourquoi ces adjectifs sont un véritable défi pour les applications informatiques multi-lingues. Dans une partie plus pratique, nous proposons une formalisation de ces adjectifs permettant de rendre compte de leurs liens avec leur base nominale. Nous tentons d'extraire ces informations dans les lexiques informatisés existants, puis nous les exploitons pour traduire les adjectifs relationnels préfixés de l'italien en français.

Face à la prolifération des publications en biologie et médecine (plus de 18 millions de publications actuellement recensées dans PubMed), l'extraction d'information automatique est devenue un enjeu crucial. Il existe de nombreux travaux dans le domaine du traitement de la langue appliquée à la biomédecine ("BioNLP"). Ces travaux se distribuent en deux grandes tendances. La première est fondée sur les méthodes d'apprentissage automatique de type numérique qui donnent de bons résultats mais ont un fonctionnement de type "boite noire". La deuxième tendance est celle du TALN à base d'analyses (lexicales, syntaxiques, voire sémantiques ou discursives) coûteuses en temps de développement des ressources nécessaires (lexiques, grammaires, etc.). Nous proposons dans cet article une approche basée sur la découverte de motifs séquentiels pour apprendre automatiquement les ressources linguistiques, en l'occurrence les patrons linguistiques qui permettent l'extraction de l'information dans les textes. Plusieurs aspects méritent d'être soulignés : cette approche permet de s'affranchir de l'analyse syntaxique de la phrase, elle ne nécessite pas de ressources en dehors du corpus d'apprentissage et elle ne demande que très peu d'intervention manuelle. Nous illustrons l'approche sur le problème de la détection d'interactions entre gènes et donnons les résultats obtenus sur des corpus biologiques qui montrent l'intérêt de ce type d'approche.

Dans un système standard de traduction statistique basé sur les segments, le score attribué aux différentes traductions d'un segment ne dépend pas du contexte dans lequel il apparaît. Plusieurs travaux récents tendent à montrer l'intérêt de prendre en compte le contexte source lors de la traduction, mais ces études portent sur des systèmes traduisant vers l'anglais, une langue faiblement fléchie. Dans cet article, nous décrivons nos expériences sur la prise en compte du contexte source dans un système statistique traduisant de l'anglais vers le français, basé sur l'approche proposée par Stroppa et al. (2007). Nous étudions l'impact de différents types d'indices capturant l'information contextuelle, dont des dépendances syntaxiques typées. Si les mesures automatiques d'évaluation de la qualité d'une traduction ne révèlent pas de gains significatifs de notre système par rapport à un système à l'état de l'art ne faisant pas usage du contexte, une évaluation manuelle conduite sur 100 phrases choisies aléatoirement est en faveur de notre système. Cette évaluation fait également ressortir que la prise en compte de certaines dépendances syntaxiques est bénéfique à notre système.

Nous assistons actuellement en TAL à un regain d'intérêt pour le traitement de la temporalité véhiculée par les textes. Dans cet article, nous présentons une proposition de caractérisation et de typage des expressions temporelles tenant compte des travaux effectués dans ce domaine tout en cherchant à pallier les manques et incomplétudes de certains de ces travaux. Nous explicitons comment nous nous situons par rapport à l'existant et les raisons pour lesquelles parfois nous nous en démarquons. Le typage que nous définissons met en évidence de réelles différences dans l'interprétation et le mode de résolution référentielle d'expressions qui, en surface, paraissent similaires ou identiques. Nous proposons un ensemble des critères objectifs et linguistiquement motivés permettant de reconnaître, de segmenter et de typer ces expressions. Nous verrons que cela ne peut se réaliser sans considérer les procès auxquels ces expressions sont associées et un contexte parfois éloigné.

L'évaluation de l'efficacité d'algorithmes de segmentation thématique est généralement effectuée en quantifiant le degré d'accord entre une segmentation hypothétique et une segmentation de référence. Les indices classiques de précision et de rappel étant peu adaptés à ce domaine, WindowDiff (Pevzner, Hearst, 2002) s'est imposé comme l'indice de référence. Une analyse de cet indice montre toutefois qu'il présente plusieurs limitations. L'objectif de ce rapport est d'évaluer un indice proposé par Bookstein, Kulyukin & Raita (2002), la distance de Hamming généralisée, qui est susceptible de remédier à celles-ci. Les analyses montrent que celui-ci conserve tous les avantages de WindowDiff sans les limitations. De plus, contrairement à WindowDiff, il présente une interprétation simple puisqu'il correspond à une vraie distance entre les deux segmentations à comparer.

Cet article vise la description et le repérage automatique des segments d'obsolescence dans les documents de type encyclopédique. Nous supposons que des indices sémantiques et discursifs peuvent permettre le repérage de tels segments. Pour ce faire, nous travaillons sur un corpus annoté manuellement par des experts sur lequel nous projetons des indices repérés automatiquement. Les techniques statistiques de base ne permettent pas d'expliquer ce phénomène complexe. Nous proposons l'utilisation de techniques de fouille de données pour le caractériser et nous évaluons le pouvoir prédictif de nos indices. Nous montrons, à l'aide de techniques de classification supervisée et de calcul de l'aire sous la courbe ROC, que nos hypothèses sont pertinentes.

Le traitement des langues fait face à une demande croissante en matière d'analyse de textes véhiculant des critiques ou des opinions. Nous présentons ici un système de résumé automatique tourné vers l'analyse d'articles postés sur des blogues, où sont exprimées à la fois des informations factuelles et des prises de position sur les faits considérés. Nous montrons qu'une approche

classique à base de traits de surface est tout à fait efficace dans ce cadre. Le système est évalué à travers une participation à la campagne d'évaluation internationale TAC (Text Analysis Conference) où notre système a réalisé des performances satisfaisantes.

Cet article décrit une méthodologie visant la réalisation d'une ressource sémantique en français centrée sur la synonymie. De manière complémentaire aux travaux existants, la méthode proposée n'a pas seulement pour objectif d'établir des liens de synonymie entre lexèmes, mais également d'apparier les sens possibles d'un lexème avec les ensembles de synonymes appropriés. En pratique, les sens possibles des lexèmes proviennent des définitions du TLFi et les synonymes de cinq dictionnaires accessibles à l'ATILF. Pour évaluer la méthode d'appariement entre sens d'un lexème et ensemble de synonymes, une ressource de référence a été réalisée pour 27 verbes du français par quatre lexicographes qui ont spécifié manuellement l'association entre verbe, sens (définition TLFi) et ensemble de synonymes. Relativement à ce standard étalon, la méthode d'appariement affiche une F-mesure de 0.706 lorsque l'ensemble des paramètres est pris en compte, notamment la distinction pronominal / non-pronominal pour les verbes du français et de 0.602 sans cette distinction.

Nous proposons un modèle filtrant de résolution de co-références basé sur les notions de transitivité et d'exclusivité linguistique. À partir de l'hypothèse générale que les chaînes de co-référence demeurent cohérentes tout au long d'un texte, notre modèle assure le respect de certaines contraintes linguistiques (via des filtres) quant à la co-référence, ce qui améliore la résolution globale. Le filtrage a lieu à différentes étapes de l'approche standard (c-à-d. par apprentissage automatique), y compris avant l'apprentissage et avant la classification, accélérant et améliorant ce processus.

La couverture d'un analyseur syntaxique dépend avant tout de la grammaire et du lexique sur lequel

il repose. Le développement d'un lexique complet et précis est une tâche ardue et de longue haleine, surtout lorsque le lexique atteint un certain niveau de qualité et de couverture. Dans cet article, nous présentons un processus capable de détecter automatiquement les entrées manquantes ou incomplètes d'un lexique, et de suggérer des corrections pour ces entrées. La détection se réalise au moyen de deux techniques reposant soit sur un modèle statistique, soit sur les informations fournies par un étiqueteur syntaxique. Les hypothèses de corrections pour les entrées lexicales détectées sont générées en étudiant les modifications qui permettent d'améliorer le taux d'analyse des phrases dans lesquelles ces entrées apparaissent. Le processus global met en oeuvre plusieurs techniques utilisant divers outils tels que des étiqueteurs et des analyseurs syntaxiques ou des classifieurs d'entropie. Son application au Lefff , un lexique morphologique et syntaxique à large couverture du français, nous a déjà permis de réaliser des améliorations notables.

Les analyseurs syntaxiques de surface à base de règles se caractérisent par un processus en deux temps : désambiguïsation lexicale, puis reconnaissance de patrons. Considérant que ces deux étapes introduisent une certaine redondance dans la description linguistique et une dilution des heuristiques dans les différents processus, nous proposons de définir un analyseur de surface qui fonctionne sur une entrée non désambiguïsée et produise l'ensemble des analyses possibles en termes de syntagmes noyau (chunks). L'analyseur, implanté avec NooJ, repose sur la définition de patrons étendus qui annotent des séquences de syntagmes noyau. Les résultats obtenus sur un corpus de développement d'environ 22 500 mots, avec un rappel proche de 100 %, montrent la faisabilité de l'approche et signalent quelques points d'ambiguïté à étudier plus particulièrement pour améliorer la précision.

Les techniques de résumé automatique multi-documents par extraction ont récemment évolué vers des méthodes statistiques pour la sélection des phrases à extraire. Dans cet article, nous

présentons un système conforme à l'« état de l'art » ? CBSEAS ? que nous avons développé pour les tâches Opinion (résumés d'opinions issues de blogs) et Update (résumés de dépêches et mise à jour du résumé à partir de nouvelles dépêches sur le même événement) de la campagne d'évaluation TAC 2008, et montrons l'intérêt d'analyses structurelles et linguistiques des documents à résumer. Nous présentons également notre étude sur la structure des dépêches et l'impact de son intégration à CBSEAS.

Nous présentons une expérience de fusion d'annotations d'entités nommées provenant de différents annotateurs. Ce travail a été réalisé dans le cadre du projet Infom@gic, projet visant à l'intégration et à la validation d'applications opérationnelles autour de l'ingénierie des connaissances et de l'analyse de l'information, et soutenu par le pôle de compétitivité Cap Digital « Image, MultiMédia et Vie Numérique ». Nous décrivons tout d'abord les quatre annotateurs d'entités nommées à l'origine de cette expérience. Chacun d'entre eux fournit des annotations d'entités conformes à une norme développée dans le cadre du projet Infom@gic. L'algorithme de fusion des annotations est ensuite présenté ; il permet de gérer la compatibilité entre annotations et de mettre en évidence les conflits, et ainsi de fournir des informations plus fiables. Nous concluons en présentant et interprétant les résultats de la fusion, obtenus sur un corpus de référence annoté manuellement.

Nous présentons une méthode de Traduction Automatique d'Unités Lexicales Complexes (ULC) pour la construction de ressources bilingues français/anglais, basée sur un système modulaire qui prend en compte les propriétés linguistiques des unités sources (compositionnalité, polysémie, etc.). Notre système exploite les différentes « facettes » du Web multi-lingue pour valider des traductions candidates ou acquérir de nouvelles traductions. Après avoir collecté une base d'ULC en français à partir d'un corpus de pages Web, nous passons par trois phases de traduction qui s'appliquent à un cas linguistique, avec une méthode adaptée : les traductions compositionnelles non polysémiques, les traductions compositionnelles polysémiques et les traductions non compositionnelles et/ou

inconnues. Notre évaluation sur un vaste échantillon d'ULC montre que l'exploitation du Web pour la traduction et la prise en compte des propriétés linguistiques au sein d'un système modulaire permet une acquisition automatique de traductions avec une excellente précision.

Un des problèmes majeurs de la linguistique aujourd'hui réside dans la prise en compte de phénomènes relevant de domaines et de modalités différentes. Dans la littérature, la réponse consiste à représenter les relations pouvant exister entre ces domaines de façon externe, en termes de relation de structure à structure, s'appuyant donc sur une description distincte de chaque domaine ou chaque modalité. Nous proposons dans cet article une approche différente permettant représenter ces phénomènes dans un cadre formel unique, permettant de rendre compte au sein d'une même grammaire tous les phénomènes concernés. Cette représentation précise de l'interaction entre domaines et modalités s'appuie sur la définition de relations d'alignement.

La tâche, aujourd'hui considérée comme fondamentale, de reconnaissance d'entités nommées, présente des difficultés spécifiques en matière d'annotation. Nous les précisons ici, en les illustrant par des expériences d'annotation manuelle dans le domaine de la microbiologie. Ces problèmes nous amènent à reposer la question fondamentale de ce que les annotateurs doivent annoter et surtout, pour quoi faire. Nous identifions pour cela les applications nécessitant l'extraction d'entités nommées et, en fonction des besoins de ces applications, nous proposons de définir sémantiquement les éléments à annoter. Nous présentons ensuite un certain nombre de recommandations méthodologiques permettant d'assurer un cadre d'annotation cohérent et évaluable.

Dans cet article, nous prenons position par rapport à la question de la qualité des données bilingues destinées à la traduction automatique statistique en terme de langue source et direction de traduction originales à l'égard d'une tâche de traduction français-anglais. Nous montrons que

l'entraînement sur un corpus contenant des textes qui ont été à l'origine traduits du français vers l'anglais améliore la qualité de la traduction. Inversement, l'entraînement sur un corpus contenant exclusivement des textes dont la langue source originale n'est ni le français ni l'anglais dégrade la traduction.

L'étape de la désambiguïsation lexicale est souvent esquivée dans les systèmes de Traduction Automatique Statistique (Statistical Machine Translation (SMT)) car considérée comme non nécessaire à la sélection de traductions correctes. Le débat autour de cette nécessité est actuellement assez vif. Dans cet article, nous présentons les principales positions sur le sujet. Nous analysons les avantages et les inconvénients de la conception actuelle de la désambiguïsation dans le cadre de la SMT, d'après laquelle les sens des mots correspondent à leurs traductions dans des corpus parallèles. Ensuite, nous présentons des arguments en faveur d'une analyse plus poussée des informations sémantiques induites à partir de corpus parallèles et nous expliquons comment les résultats d'une telle analyse pourraient être exploités pour une évaluation plus flexible et concluante de l'impact de la désambiguïsation dans la SMT.

L'évaluation des systèmes de dialogue homme-machine est un problème difficile et pour lequel ni les objectifs ni les solutions proposées ne font aujourd'hui l'unanimité. Les approches ergonomiques traditionnelles soumettent le système de dialogue au regard critique de l'utilisateur et tente d'en capter l'expression, mais l'absence d'un cadre objectivable des usages de ces utilisateurs empêche une comparaison entre systèmes différents, ou entre évolutions d'un même système. Nous proposons d'inverser cette vision et de mesurer le comportement de l'utilisateur au regard du système de dialogue. Aussi, au lieu d'évaluer l'adéquation du système à ses utilisateurs, nous mesurons l'adéquation des utilisateurs au système. Ce changement de paradigme permet un changement de référentiel qui n'est plus les usages des utilisateurs mais le cadre du système. Puisque le système est complètement défini, ce paradigme permet des approches quantitatives et

donc des évaluations comparatives de systèmes.

Dans une tâche consistant à trouver l'auteur (parmi 53) de chacun de 114 textes, nous analysons la performance de modèles de langue et de modèles stylométriques sous les angles du rappel et du nombre de paramètres. Le modèle de mots bigramme à lissage de Kneser-Ney modifié interpolé est le plus performant (75 % de bonnes réponses au premier rang). Parmi les modèles stylométriques, une combinaison de 7 paramètres liés aux parties du discours produit les meilleurs résultats (rappel de 25 % au premier rang). Dans les deux catégories de modèles, le rappel maximal n'est pas atteint lorsque le nombre de paramètres est le plus élevé.

La densité des idées, qui correspond au ratio entre le nombre de propositions sémantiques et le nombre de mots dans un texte reflète la qualité informative des propositions langagières d'un texte. L'apparition de la maladie d'Alzheimer a été reliée à une dégradation de la densité des idées, ce qui explique l'intérêt pour un calcul automatique de cette mesure. Nous proposons une méthode basée sur un étiquetage morpho-syntaxique et des règles d'ajustement, inspirée du logiciel CPIDR. Cette méthode a été validée sur un corpus de quarante entretiens oraux transcrits et obtient de meilleurs résultats pour le français que CPIDR pour l'anglais. Elle est implémentée dans le logiciel libre Densidées disponible sur http://code.google.com/p/densidees.

La segmentation en mots est une première étape possible dans le traitement automatique de la langue chinoise. Les systèmes de segmentation se sont beaucoup développés depuis le premier apparu dans les années 1980. Il n'existe cependant aucun outil standard aujourd'hui. L'objectif de ce travail est de faire une comparaison des différents outils de segmentation en s'appuyant sur une analyse statistique. Le but est de définir pour quel type de texte chacun d'eux est le plus performant. Quatre outils de segmentation et deux corpus avec des thèmes distincts ont été choisis pour cette étude. À l'aide des outils textométriques Lexico3 et mkAlign, nous avons centré notre analyse sur le

nombre de syllabes du chinois. Les données quantitatives ont permis d'objectiver des différences entre les outils. Le système Hylanda s'avère performant dans la segmentation des termes spécialisés et le système Stanford est plus indiqué pour les textes généraux. L'étude de la comparaison des outils de segmentation montre le statut incontournable de l'analyse textométrique aujourd'hui, celle-ci permettant d'avoir accès rapidement à la recherche d'information.

La constitution de ressources linguistiques est une tâche cruciale pour les systèmes d'extraction d'information fondés sur une approche symbolique. Ces systèmes reposent en effet sur des grammaires utilisant des informations issues de dictionnaires électroniques ou de réseaux sémantiques afin de décrire un phénomène linguistique précis à rechercher dans les textes. La création et la révision manuelle de telles ressources sont des tâches longues et coûteuses en milieu industriel. Nous présentons ici un nouvel algorithme produisant une grammaire d'extraction de relations entre entités nommées, de manière semi-automatique à partir d'un petit ensemble de phrases représentatives. Dans un premier temps, le linguiste repère un jeu de phrases pertinentes à partir d'une analyse des co-occurrences d'entités repérées automatiquement. Cet échantillon n'a pas forcément une taille importante. Puis, un algorithme permet de produire une grammaire en généralisant progressivement les éléments lexicaux exprimant la relation entre entités. L'originalité de l'approche repose sur trois aspects : une représentation riche du document initial permettant des généralisations pertinentes, la collaboration étroite entre les aspects automatiques et l'apport du linguiste et sur la volonté de contrôler le processus en ayant toujours affaire à des données lisibles par un humain.

Les corpus de paraphrases à large échelle sont importants dans de nombreuses applications de TAL. Dans cet article nous présentons une méthode visant à obtenir un corpus parallèle de paraphrases d'énoncés en français. Elle vise à collecter des traductions multiples proposées par des contributeurs volontaires francophones à partir de plusieurs langues européennes. Nous

formulons l'hypothèse que deux traductions soumises indépendamment par deux participants conservent généralement le sens de la phrase d'origine, quelle que soit la langue à partir de laquelle la traduction est effectuée. L'analyse des résultats nous permet de discuter cette hypothèse.

La détection des informations temporelles est cruciale pour le traitement automatique des textes, qu'il s'agisse de modélisation linguistique, d'applications en compréhension du langage ou encore de tâches de recherche documentaire ou d'extraction d'informations. De nombreux travaux ont été dédiés à l'analyse temporelle des textes, et plus précisément l'annotation des expressions temporelles ou des événements sous leurs différentes formes : verbales, adjectivales ou nominales. Dans cet article, nous décrivons une méthode pour la détection des syntagmes nominaux dénotant des événements. Notre approche est basée sur l'implémentation d'un test linguistique simple proposé par les linguistes pour cette tâche. Nous avons expérimenté notre méthode sur deux corpus différents ; le premier est composé d'articles de presse et le second est beaucoup plus grand, utilisant une interface pour interroger automatiquement le moteur de recherche Yahoo. Les résultats obtenus ont montré que cette méthode se révèle plus pertinente pour un plus large corpus.

Cet article décrit un processus d'annotation manuelle de textes d'opinion, basé sur un schéma fin d'annotation indépendant de la langue et du corpus. Ensuite, à partir d'une partie de ce schéma, une méthode de construction automatique d'un lexique d'opinion à partir d'un analyseur syntaxique et d'une ressource linguistique est décrite. Cette méthode consiste à construire un arbre de décision basé sur les classes de concepts de la ressource utilisée. Dans un premier temps, nous avons étudié la couverture du lexique d'opinion obtenu par comparaison avec l'annotation manuelle effectuée sur un premier corpus de critiques de restaurants. La généricité de ce lexique a été mesurée en le comparant avec un second lexique, généré à partir d'un corpus de commentaires de films. Dans un second temps, nous avons évalué l'utilisabilité du lexique au travers d'une tâche extrinsèque, la reconnaissance de la polarité de commentaires d'internautes.

Dans cet article, nous nous intéressons au résumé automatique de textes arabes. Nous commençons par présenter une étude analytique réalisée sur un corpus de travail qui nous a permis de déduire, suite à des observations empiriques, un ensemble de relations et de frames (règles ou patrons) rhétoriques; ensuite nous présentons notre méthode de production de résumés pour les textes arabes. La méthode que nous proposons se base sur la Théorie de la Structure Rhétorique (RST) (Mann et al., 1988) et utilise des connaissances purement linguistiques. Le principe de notre proposition s'appuie sur trois piliers. Le premier pilier est le repérage des relations rhétoriques entres les différentes unités minimales du texte dont l'une possède le statut de noyau ? segment de texte primordial pour la cohérence ? et l'autre a le statut noyau ou satellite ? segment optionnel. Le deuxième pilier est le dressage et la simplification de l'arbre RST. Le troisième pilier est la sélection des phrases noyaux formant le résumé final, qui tiennent en compte le type de relation rhétoriques choisi pour l'extrait.

Ce travail s'inscrit dans une recherche centrée sur une approche de l'Intelligence Artificielle (IA) et de la linguistique computationnelle. Il permet d'intégrer différentes techniques formelles de la Logique Combinatoire avec des types (Curry) et sa programmation fonctionnelle (Haskell) avec une théorie énonciative du temps et de l'aspect. Nous proposons des calculs formels de valeurs aspecto-temporelles (processus inaccompli présent, processus inaccompli passé, événement passé et étatrésultant présent) associées à des représentations de significations verbales sous forme de schèmes applicatifs.

Dans cet article, nous présentons une analyse manuelle de corpus de contextes conceptuels afin (i) de voir dans quelle mesure les méthodes de TALN existantes sont en principe adéquates pour automatiser la rédaction de définitions terminographiques, et (ii) de dégager des question précises dont la résolution permettrait d'automatiser davantage la production de définitions. Le but est de

contribuer à la réflexion sur les enjeux de l'automatisation de cette tâche, en procédant à une série d'analyses qui nous mènent, étape par étape, à examiner l'adéquation des méthodes d'extraction de définitions et de contextes plus larges au travail terminographique de rédaction des définitions. De ces analyses émergent des questions précises relatives à la pertinence des informations extraites et à leur sélection. Des propositions de solutions et leurs implications pour le TALN sont examinées.

Avec le développement d'internet et des sites d'échanges (forums, blogs, sondages en ligne, ...), l'exploitation de nouvelles sources d'informations dans le but d'en extraire des opinions sur des sujets précis (film, commerce,...) devient possible. Dans ce papier, nous présentons une approche de fouille d'opinions à partir de textes courts. Nous expliquons notamment en quoi notre choix d'utilisation de regroupements autour des idées exprimées nous a conduit à opter pour une représentation implicite telle que la représentation vectorielle. Nous voyons également les différents traitements sémantiques intégrés à notre chaîne de traitement (traitement de la négation, lemmatisation, stemmatisation, synonymie ou même polysémie des mots) et discutons leur impact sur la qualité des regroupements obtenus.

Cet article présente une étude en corpus comparable médical pour confirmer la préférence d'utilisation des adjectifs relationnels dans les langues de spécialité et examiner plus finement l'alternance entre syntagmes nominaux avec adjectifs relationnels et syntagmes avec complément prépositionnel.

Grâce à la participation d'un grand nombre de personnes via des jeux accessibles sur le web, nous avons construit un réseau lexical évolutif de grande taille pour le Français. A partir de cette ressource, nous avons abordé la question de la détermination des sens d'usage d'un terme, puis après avoir introduit la notion de similarité entre ces différents usages, nous avons pu obtenir pour un terme son arbre des usages : la racine regroupe tous les usages du terme et une descente dans

l'arbre correspond à un raffinement de ces usages. Le nommage des différents noeuds est effectué lors d'une descente en largeur. En simplifiant l'arbre des usages nommés, nous déterminons les différents sens d'un terme, sens que nous introduisons dans le réseau lexical en tant que noeuds de raffinement du terme considéré. Nous terminons par une évaluation empirique des résultats obtenus.

Les définitions des paraphrases privilégient généralement la conservation du sens. Cet article démontre par l'absurde qu'une évaluation uniquement basée sur la conservation du sens permet à un système inutile de production de paraphrase d'être jugé meilleur qu'un système au niveau de l'état de l'art. La conservation du sens n'est donc pas l'unique critère des paraphrases. Nous exhibons les trois objectifs des paraphrases : la conservation du sens, la naturalité et l'adaptation à la tâche. La production de paraphrase est alors un compromis dépendant de la tâche entre ces trois critères et ceux-ci doivent être pris en compte lors des évaluations.

Le travail présenté dans cet article s'inscrit dans le thème de l'acquisition automatique de ressources sémantiques s'appuyant sur les données de Wikipedia. Nous exploitons le graphe des catégories associées aux pages de Wikipedia à partir duquel nous extrayons une hiérarchie de catégories parentes, sémantiquement et thématiquement liées. Cette extraction est le résultat d'une stratégie de plus court chemin appliquée au treillis global des catégories. Chaque page peut ainsi être représentée dans l'espace de ses catégories propres, ainsi que des catégories parentes. Nous montrons la possibilité d'utiliser cette ressource pour deux applications. La première concerne l'indexation et la classification des pages de Wikipedia. La seconde concerne la désambiguïsation dans le cadre d'un traducteur de requêtes français/anglais. Ce dernier travail a été réalisé en exploitant les catégories des pages anglaises.

Certaines ponctuations fortes sont « abusivement » utilisées à la place de ponctuations faibles,

débouchant sur des phrases graphiques qui ne sont pas des phrases grammaticales. Cet article présente une étude sur corpus de ce phénomène et une ébauche d'outil pour repérer automatiquement les ponctuations fortes abusives.

La plupart des systèmes de question-réponse ont été conçus pour répondre à des questions dites "factuelles" (réponses précises comme des dates, des lieux), et peu se sont intéressés au traitement des questions complexes. Cet article présente une typologie des questions en y incluant les questions complexes, ainsi qu'une typologie des formes de réponses attendues pour chaque type de questions. Nous présentons également des expériences préliminaires utilisant ces typologies pour les questions complexes, avec de bons résultats.

La TA généraliste de haute qualité et totalement automatique est considérée comme impossible. Nous nous intéressons aux problèmes de traduction scripturale, qui sont des sous-problèmes faibles du problème général de la traduction. Nous présentons les caractéristiques des problèmes faibles de traduction et les problèmes de traduction scripturale, décrivons différentes approches computationnelles (à états finis, statistiques, et hybrides) et présentons nos résultats sur différentes combinaisons de langues et systèmes d'écriture Indo-Pak.

Dans le cadre de la promotion de la langue amazighe, nous avons voulu lui apporter des ressources et outils linguistiques pour son traitement automatique et son intégration dans le domaine des nouvelles technologies de l'information et de la communication. Partant de ce principe, nous avons opté, au sein de l'Institut Royal de la Culture Amazighe, pour une démarche innovante de réalisations progressives de ressources linguistiques et d'outils de base de traitement automatique, qui permettront de préparer le terrain pour d'éventuelles recherches scientifiques. Dans cette perspective, nous avons entrepris de développer, dans un premier temps, un outil de pseudoracinisation basé sur une approche relevant du cas de la morphologie flexionnelle et

reposant sur l'élimination d'une liste de suffixes et de préfixes de la langue amazighe. Cette approche permettra de regrouper les mots sémantiquement proches à partir de ressemblances afin d'être exploités dans des applications tel que la recherche d'information et la classification.

Cet article montre comment calculer une interface syntaxe-sémantique à partir d'un analyseur en dépendance quelconque et interchangeable, de ressources lexicales variées et d'une base d'exemples associés à leur représentation sémantique. Chaque exemple permet de construire une règle d'interface. Nos représentations sémantiques sont des graphes hiérarchisés de relations prédicat-argument entre des acceptions lexicales et notre interface syntaxe-sémantique est une grammaire de correspondance polarisée. Nous montrons comment obtenir un système très modulaire en calculant certaines règles par « soustraction » de règles moins modulaires.

Dans cet article, nous abordons la problématique de la traduction automatique des pronoms clitiques, en nous focalisant sur la traduction de l'italien vers le français et en comparant les résultats obtenus par trois systèmes : Its-2, développé au LATL (Laboratoire d'Analyse et de Technologie du Langage) et basé sur un analyseur syntaxique profond ; Babelfish, basé sur des règles linguistiques ; et Google Translate, caractérisé par une approche statistique.

Cette étude utilise des outils de TAL pour tester l'hypothèse avancée par plusieurs études linguistiques récentes selon laquelle la relation ant-onymique, classiquement décrite comme une relation paradigmatique, a la particularité de fonctionner également sur le plan syntagmatique, c'est-à-dire de réunir des mots qui sont non seulement substituables mais qui apparaissent également régulièrement dans des relations contextuelles. Nous utilisons deux méthodes ? l'analyse distributionnelle pour le plan paradigmatique, la recherche par patrons ant-onymiques pour le plan syntagmatique. Les résultats montrent que le diagnostic d'antonymie n'est pas significativement meilleur lorsqu'on croise les deux méthodes, puisqu'une partie des antonymes

identifiés ne répondent pas au test de substituabilité, ce qui semble confirmer la prépondérance du plan syntagmatique pour l'étude et l'acquisition de cette relation.

EDF utilise les techniques de Text Mining pour optimiser sa relation client, en analysant des réponses aux questions ouvertes d'enquête de satisfaction, et des retranscriptions de conversations issues des centres d'appels. Dans cet article, nous présentons les différentes contraintes applicatives liées à l'utilisation d'outils de text mining pour l'analyse de données clients. Après une analyse des différents outils présents sur le marché, nous avons identifié la technologie Skill CartridgeTM fournie par la société TEMIS comme la plus adaptée à nos besoins. Cette technologie nous permet une modélisation sémantique de concepts liés au motif d'insatisfaction. L'apport de cette modélisation est illustrée pour une tâche de classification de réponses d'enquêtes de satisfaction chargée d'évaluer la fidélité des clients EDF. La modélisation sémantique a permis une nette amélioration des scores de classification (F-mesure = 75,5%) notamment pour les catégories correspondant à la satisfaction et au mécontentement.

L'étiquetage des rôles grammaticaux est une tâche de pré-traitement récurrente. Pour le français, deux outils sont majoritairement utilisés : TreeTagger et Brill. Nous proposons une démarche, ne nécessitant aucune ressource, pour la création d'un modèle de Markov caché (HMM) pour palier les problèmes de ces outils, et de licences notamment. Nous distribuons librement toutes les ressources liées à ce travail.

Cet article présente une approche basée sur la comparaison fréquentielle de modèles lexicaux pour la segmentation automatique de textes historiques Portugais. Cette approche traite d'abord le problème de la segmentation comme un problème de classification, en attribuant à chaque élément lexical présent dans la phase d'apprentissage une valeur de saillance pour chaque type de segment. Ces modèles lexicaux permettent à la fois de produire une segmentation et de faire une

analyse qualitative de textes historiques. Notre évaluation montre que l'approche adoptée permet de tirer de l'information sémantique que des approches se concentrant sur la détection des frontières séparant les segments ne peuvent acquérir.

Cet article aborde le phénomène de l'incomplétude des ressources lexicales, c'est-à-dire la problématique des inconnus, dans un contexte de traitement automatique. Nous proposons tout d'abord une définition opérationnelle de la notion d'inconnu. Nous décrivons ensuite une typologie des différentes classes d'inconnus, motivée par des considérations linguistiques et applicatives ainsi que par l'annotation des inconnus d'un petit corpus selon notre typologie. Cette typologie sera mise en oeuvre et validée par l'annotation d'un corpus important de l'Agence France-Presse dans le cadre du projet EDyLex.

Cet article s'inscrit dans les recherches sur l'exploitation de ressources terminologiques pour l'analyse de textes de spécialité, leur annotation et leur indexation. Les ressources en présence sont, d'une part, un thesaurus des Sciences du Langage, le Thesaulangue et, d'autre part, un corpus d'échantillons issus de cinq ouvrages relevant du même domaine. L'article a deux objectifs. Le premier est de déterminer dans quelle mesure les termes de Thesaulangue sont représentés dans les textes. Le second est d'évaluer si les occurrences des unités lexicales correspondant aux termes de Thesaulangue relèvent majoritairement d'emplois terminologiques ou de langue courante. A cette fin, les travaux présentés utilisent une mesure de richesse lexicale telle qu'elle a été définie par Brunet (rapporté dans Muller, 1992) dans le domaine de la lexicométrie, l'indice W. Cette mesure est adaptée afin de mesurer la richesse terminologie (co-occurrents lexicaux et sémantiques qui apparaissent dans Thesaulangue).

Le projet EmotiRob, soutenu par l'agence nationale de la recherche, s'est donné pour objectif de détecter des émotions dans un contexte d'application original : la réalisation d'un robot compagnon

émotionnel pour des enfants fragilisés. Nous présentons dans cet article le système qui caractérise l'émotion induite par le contenu linguistique des propos de l'enfant. Il se base sur un principe de compositionnalité des émotions, avec une valeur émotionnelle fixe attribuée aux mots lexicaux, tandis que les verbes et les adjectifs agissent comme des fonctions dont le résultat dépend de la valeur émotionnelle de leurs arguments. L'article présente la méthode de calcul utilisée, ainsi que la norme lexicale émotionnelle correspondante. Une analyse quantitative et qualitative des premières expérimentations présente les différences entre les sorties du module de détection et l'annotation d'experts, montrant des résultats satisfaisants, avec la bonne détection de la valence émotionnelle dans plus de 90% des cas.

Dans cet article, nous présentons la tâche d'acquisition de paraphrases sous-phrastiques (impliquant des paires de mots ou de groupes de mots), et décrivons plusieurs techniques opérant à différents niveaux. Nous décrivons une évaluation visant à comparer ces techniques et leurs combinaisons sur deux corpus de paraphrases d'énoncés obtenus par traduction multiple. Les conclusions que nous tirons peuvent servir de guide pour améliorer des techniques existantes.

Word-Net, une des ressources lexicales les plus utilisées aujourd'hui a été constituée en anglais et les chercheurs travaillant sur d'autres langues souffrent du manque d'une telle ressource. Malgré les efforts fournis par la communauté française, les différents Word-Nets produits pour la langue française ne sont toujours pas aussi exhaustifs que le Word-Net de Princet-on. C'est pourquoi nous proposons une méthode novatrice dans la production de termes nominaux instanciant les différents synsets de Word-Net en exploitant les propriétés syntaxiques distributionnelles du vocabulaire français. Nous comparons la ressource que nous obtenons avecWOLF et montrons que notre approche offre une couverture plus large.

Dans cet article nous relatons notre participation à la campagne d'évaluation ESTER 2 (Evaluation

des Systèmes de Transcription Enrichie d'Emissions Radiophoniques). Après avoir décrit les objectifs de cette campagne ainsi que ses spécificités et difficultés, nous présentons notre système d'extraction d'entités nommées en nous focalisant sur les adaptations réalisées dans le cadre de cette campagne. Nous décrivons ensuite les résultats obtenus lors de la compétition, ainsi que des résultats originaux obtenus par la suite. Nous concluons sur les leçons tirées de cette expérience.

Dans cet article, nous présentons RefGen, un module d'identification des chaînes de référence pour le français. RefGen effectue une annotation automatique des expressions référentielles puis identifie les relations de co-référence établies entre ces expressions pour former des chaînes de référence. Le calcul de la référence utilise des propriétés des chaînes de référence dépendantes du genre textuel, l'échelle d'accessibilité d'(Ariel, 1990) et une série de filtres lexicaux, morpho-syntaxiques et sémantiques. Nous évaluons les premiers résultats de RefGen sur un corpus issu de rapports publics.

Nous présentons NP, un système de reconnaissance d'entités nommées. Comprenant un module de résolution, il permet d'associer à chaque occurrence d'entité le référent qu'elle désigne parmi les entrées d'un référentiel dédié. NP apporte ainsi des informations pertinentes pour l'exploitation de l'extraction d'entités nommées en contexte applicatif. Ce système fait l'objet d'une évaluation grâce au développement d'un corpus annoté manuellement et adapté aux tâches de détection et de résolution.

Cet article présente un processus de décision basé sur des classifieurs à vaste marge (SVMDP) pour extraire l'information sémantique dans un système de dialogue oral. Dans notre composant de compréhension, l'information est représentée par des arbres de frames sémantiques définies selon le paradigme FrameNet. Le processus d'interprétation est réalisé en deux étapes. D'abord, des réseaux bayésiens dynamiques (DBN) sont utilisés comme modèles de génération pour inférer des

fragments d'arbres de la requête utilisateur. Ensuite, notre SVMDP dépendant du contexte compose ces fragments afin d'obtenir la représentation sémantique globale du message. Les expériences sont menées sur le corpus de dialogue MEDIA. Une procédure semi-automatique fournit une annotation de référence en frames sur laquelle les paramètres des DBN et SVMDP sont appris. Les résultats montrent que la méthode permet d'améliorer les performances d'identification de frames pour les exemples de test les plus complexes par rapport à un processus de décision déterministe ad hoc.

L'extraction des événements désignés par des noms est peu étudiée dans des corpus généralistes. Si des lexiques de noms déclencheurs d'événements existent, les problèmes de polysémie sont nombreux et beaucoup d'événements ne sont pas introduits par des déclencheurs. Nous nous intéressons dans cet article à une hypothèse selon laquelle les verbes induisant la cause ou la conséquence sont de bons indices quant à la présence d'événements nominaux dans leur co-texte.

Un lexique affectif est un outil utile pour l'étude des émotions ainsi que pour la fouille d'opinion et l'analyse des sentiments. Un tel lexique contient des listes de mots annotés avec leurs évaluations émotionnelles. Il existe un certain nombre de lexiques affectifs pour la langue anglaise, espagnole, allemande, mais très peu pour le français. Un travail de longue haleine est nécessaire pour construire et enrichir un lexique affectif. Nous proposons d'utiliser Twitter, la plateforme la plus populaire de microblogging de nos jours, pour recueillir un corpus de textes émotionnels en français. En utilisant l'ensemble des données recueillies, nous avons estimé les normes affectives de chaque mot. Nous utilisons les données de la Norme Affective desMots Anglais (ANEW, Affective Norms of EnglishWords) que nous avons traduite en français afin de valider nos résultats. Les valeurs du coefficient tau de Kendall et du coefficient de corrélation de rang de Spearman montrent que nos scores estimés sont en accord avec les scores ANEW.

Notre travail concerne l'analyse automatique des énoncés d'opinion en chinois. En nous inspirant de la théorie linguistique de l'Appraisal, nous proposons une méthode fondée sur l'usage de lexiques et de règles locales pour déterminer les caractéristiques telles que la Force (intensité), le Focus (prototypicalité) et la polarité de tels énoncés. Nous présentons le modèle et sa mise en oeuvre sur un corpus journalistique. Si pour la détection d'énoncés d'opinion, la précision est bonne (94 %), le taux de rappel (67 %) pose cependant des questions sur l'enrichissement des ressources actuelles.

Dans cet article nous présentons une méthode utilisant l'extraction de motifs séquentiels d'itemsets pour l'apprentissage automatique de patrons linguistiques. De plus, nous proposons de nous appuyer sur l'ordre partiel existant entre les motifs pour les énumérer de façon structurée et ainsi faciliter leur validation en tant que patrons linguistiques.

En intelligence artificielle, l'analogie est utilisée comme une technique de raisonnement non exact pour la résolution de problèmes, la compréhension du langage naturel, l'apprentissage des règles de classification, etc. Cet article s'intéresse à la proportion analogique, une forme simple du raisonnement par analogie, et présente son application en apprentissage automatique pour le TALN. La proportion analogique est une relation entre quatre objets qui exprime que la manière de transformer le premier objet en le second est la même que la façon de transformer le troisième en le quatrième. Premièrement, nous définissons formellement la proportion analogique entre quatre objets. Nous nous intéressons particulièrement aux objets structurés que sont les arbres ordonnés et étiquetés, avec une définition originale de l'analogie fondée sur l'alignement optimal. Ensuite, nous présentons deux algorithmes qui calculent la dissemblance analogique entre quatre arbres et qui trouvent des solutions, éventuellement approchées, à une équation analogique entre arbres. Nous montrons leur utilisation dans deux applications : l'apprentissage de l'arbre syntaxique d'une phrase et la génération de la prosodie dans la synthèse de parole.

Dans cet article, nous présentons Esculape, un système de question-réponse en français dédié aux médecins généralistes et élaboré à partir d'OEdipe, un système de question-réponse en domaine ouvert. Esculape ajoute à OEdipe la capacité d'exploiter la structure d'un modèle du domaine, le domaine médical dans le cas présent. Malgré l'existence d'un grand nombre de ressources dans ce domaine (UMLS, MeSH ...), il n'est pas possible de se reposer entièrement sur ces ressources, et plus spécifiquement sur les relations qu'elles abritent, pour répondre aux questions. Nous montrons comment surmonter cette difficulté en apprenant de façon supervisée des patrons linguistiques d'extraction de relations et en les appliquant à l'extraction de réponses.

Dans cet article, nous proposons une méthode hybride pour la reconnaissance des entités nommées pour la langue arabe. Cette méthode profite, d'une part, des avantages de l'utilisation d'une méthode d'apprentissage pour extraire des règles permettant l'identification et la classification des entités nommées. D'autre part, elle repose sur un ensemble de règles extraites manuellement pour corriger et améliorer le résultat de la méthode d'apprentissage. Les résultats de l'évaluation de la méthode proposée sont encourageants. Nous avons obtenu un taux global de F-mesure égal à 79.24%.

Cet article décrit le développement d'une grammaire catégorielle à large couverture du Français, extraite à partir du corpus arboré de Paris 7 et vérifiée et corrigée manuellement. Le grammaire catégorielle résultant est évaluée en utilisant un supertagger et obtient des résultats comparables aux meilleurs supertaggers pour l'Anglais.

Le présent article décrit un travail en cours sur l'acquisition des patrons de relations entre entités nommées à partir de résultats d'analyse syntaxique. Sans aucun patron prédéfini, notre méthode fournit des chemins syntaxiques susceptibles de représenter une relation donnée à partir de quelques exemples de couples d'entités nommées entretenant la relation en question.

Dans cet article, nous présentons et analysons les résultats du système de reconnaissance d'entités nommées CasEN lors de sa participation à la campagne d'évaluation Ester2. Nous identifions quelles ont été les difficultés pour notre système, essentiellement : les mots hors-vocabulaire, la métonymie, les frontières des entités nommées. Puis nous proposons une approche pour améliorer les performances de systèmes à base de connaissances, en utilisant des techniques exhaustives de fouille de données séquentielles afin d'extraire des motifs qui représentent les structures linguistiques en jeu lors de la reconnaissance d'entités nommées. Enfin, nous décrivons l'expérimentation menée à cet effet, donnons les résultats obtenus à ce jour et en faisons une première analyse.

Cet article s'inscrit dans le domaine de la recherche d'information multi-lingue. Il propose une méthode de traduction automatique de requêtes basée sur Wikipédia. Une phase d'analyse permet de segmenter la requête en syntagmes ou unités lexicales à traduire en s'appuyant sur les liens multi-lingues entre les articles de Wikipédia. Une deuxième phase permet de choisir, parmi les traductions possibles, celle qui est la plus cohérente en s'appuyant sur les informations d'ordre sémantique fournies par les catégories associées à chacun des articles de Wikipédia. Cet article justifie que les données issues de Wikipédia sont particulièrement pertinentes pour la traduction de requêtes, détaille l'approche proposée et son implémentation, et en démontre le potentiel par la comparaison du taux d'erreur du prototype de traduction avec celui d'autres services de traduction automatique.

Notre démonstration porte sur le prototype TerminoWeb, une plateforme Web qui permet (1) la construction automatique d'un corpus thématique à partir d'une recherche de documents sur le Web, (2) l'extraction de termes du corpus, et (3) la recherche d'information définitionnelle sur ces termes en corpus. La plateforme intégrant les trois modules, elle aidera un langagier (terminologue,

traducteur, rédacteur) à découvrir un nouveau domaine (thème) en facilitant la recherche et l'analyse de documents informatifs pertinents à ce domaine.

Nous présenterons des iMAG (interactive multilingual Access Gateways en particulier sur un site Web d'un
laboratoire scientifique et sur le site Web de l'agglomération grenobloise (La Métro). Cette présentation
bilingue a été obtenue en utilisant une iMAG.

Dans cette démonstration, nous présentons l'implémentation d'un outil de repérage d'entités nommées à base de règle pour la langue arabe dans le système de veille médiatique EMM (Europe Media Monitor).

Nous présentons Expressive, un système de génération de parole expressive à partir de données non linguistiques. Ce système est composé de deux outils distincts : Taittingen, un générateur automatique de textes d'une grande variété lexico-syntaxique produits à partir d'une représentation conceptuelle du discours, et StyloPhone, un système de synthèse vocale multi-styles qui s'attache à rendre le discours produit attractif et naturel en proposant différents styles vocaux.

Cet article présente une approche et des résultats utilisant l'encyclopédie en ligne Wikipédia comme ressource semi-structurée de connaissances linguistiques et en particulier comme un corpus comparable pour l'extraction de terminologie bilingue. Cette approche tend à extraire d'abord des paires de terme et traduction à partir de types des informations, liens et textes de Wikipédia. L'étape suivante consiste à l'utilisation de l'information linguistique afin de ré-ordonner les termes et leurs traductions pertinentes et ainsi éliminer les termes cibles inutiles. Les évaluations préliminaires utilisant les paires de langues français-anglais, japonais-français et japonais-anglais ont montré une

bonne qualité des paires de termes extraits. Cette étude est très favorable pour la construction et l'enrichissement des ressources linguistiques tels que les dictionnaires et ontologies multi-lingues. Aussi, elle est très utile pour un système de recherche d'information translinguistique (RIT).

Cet article présente Dicta-Sign, un projet de recherche sur le traitement automatique des langues des signes (LS), qui aborde un grand nombre de questions de recherche : linguistique de corpus, modélisation linguistique, reconnaissance et génération automatique. L'objectif de ce projet est de réaliser trois applications prototypes destinées aux usagers sourds : un traducteur de termes de LS à LS, un outil de recherche par l'exemple et un Wiki en LS. Pour cela, quatre corpus comparables de cinq heures de dialogue seront produits et analysés. De plus, des avancées significatives sont attendues dans le domaine des outils d'annotation. Dans ce projet, le LIMSI est en charge de l'élaboration des modèles linguistiques et participe aux aspects corpus et génération automatique. Nous nous proposons d'illustrer l'état d'avancement de Dicta-Sign au travers de vidéos extraites du corpus et de démonstrations des outils de traitement et de génération d'animations de signeur virtuel.

L'analyseur multi-lingue FiPS permet de transformer une phrase en une structure syntaxique riche et accompagnée d'informations lexicales, grammaticales et thématiques. On décrit ici une application qui adapte les structures en constituants de l'analyseur FiPS à une nomenclature grammaticale permettant la représentation en couleur. Cette application interactive et disponible en ligne (http://latl.unige.ch/fipscolor) peut être utilisée librement par les enseignants et élèves de primaire.

Cette démonstration présente une interface web pour des données numérisées de l'atlas linguistique de la Suisse allemande. Nous présentons d'abord l'intégration des données brutes et des données interpolées de l'atlas dans une interface basée sur Google Maps. Ensuite, nous

montrons des prototypes de systèmes de traduction automatique et d'identification de dialectes qui s'appuient sur ces données dialectologiques numérisées.

Cet article présente Text-it / Voice-it, une application de normalisation des SMS pour téléphone mobile. L'application permet d'envoyer et de recevoir des SMS normalisés, et offre le choix entre un résultat textuel (Text-it) et vocal (Voice-it).

Cette démonstration décrit Grail : un analyseur syntaxique pour grammaires catégorielles. Elle met l'accent sur les recherches récentes qui ont permis à Grail de donner des analyses syntaxiques et sémantiques du Français. Ces développements sont possibles grâce à une grammaire extraite semi-automatiquement du corpus de Paris 7 ainsi qu'un lexique sémantique qui traduit des combinaisons de mots, des étiquettes syntaxiques et des formules en Discourse Representation Structures.

Nous présentons une plate-forme d'annotation sémantique et d'exploration de textes médicaux, appelée « MeTAE ». Le processus d'annotation automatique comporte une première étape de reconnaissance des entités médicales présentes dans les textes suivie d'une étape d'identification des relations sémantiques qui les relient. Cette identification se fonde sur des patrons linguistiques construits manuellement pour chaque type de relation. MeTAE génère des annotations RDF à partir des informations extraites et offre une interface d'exploration des textes annotés avec des requêtes sous forme de formulaire. La plate-forme peut être utilisée pour analyser sémantiquement les textes médicaux ou interroger la base d'annotation disponible pour avoir une/des réponses à une requête donnée (e.g. « ?X prévient maladie d'Alzheimer », équivalent à la question « comment prévenir la maladie d'Alzheimer ? »). Cette application peut être la base d'un système de questions-réponses pour le domaine médical.

Nous présentons ici l'analyseur syntaxique LEOPAR basé sur les grammaires d'interaction ainsi que d'autres outils utiles pour notre chaîne de traitement syntaxique.

Malgré les nombreuses études visant à améliorer la traduction automatique, la traduction assistée par ordinateur reste la solution préférée des traducteurs lorsqu'une sortie de qualité est recherchée. Cette démonstration vise à présenter le moteur de recherche de traductions TransSearch. Cetteapplication commerciale, accessible sur leWeb, repose d'une part sur l'exploitation d'un bi-texte aligné au niveau des phrases, et d'autre part sur des modèles statistiques d'alignement de mots.

Description de Moz, un système d'aide à la traduction conçu pour le traitement de textes structurés ou semi-structurés avec une forte proportion de contenu terminologique. Le système comporte une mémoire de traduction collaborative, qui atteint un niveau élevé de rappel grâce à l'analyse sous-phrastique ; il fournit également des dispositifs de communication et de révision. Le système est en production et traduit 140 000 mots par semaine.

Les performances d'un système de traduction statistique dépendent beaucoup de la qualité et de la quantité des données d'apprentissage disponibles. La plupart des textes parallèles librement disponibles proviennent d'organisations internationales. Le jargon observé dans ces textes n'est pas très adapté pour construire un système de traduction pour d'autres domaines. Nous présentons dans cet article une technique pour adapter le modèle de traduction à un domaine différent en utilisant des textes dans la langue source uniquement. Nous obtenons des améliorations significatives du score BLEU dans des systèmes de traduction de l'arabe vers le français et vers l'anglais.

Bien souvent, le sens d'un mot ou d'une expression peut être rendu dans une autre langue par

plusieurs traductions. Parmi celles-ci, certaines se révèlent très fréquentes alors que d'autres le sont beaucoup moins, conformément à une loi zipfienne. La googlisation de notre monde n'échappe pas aux mémoires de traduction, qui mettent souvent à mal ou simplement ignorent ces traductions rares qui sont souvent de bonne qualité. Dans cet article, nous nous intéressons à ces traductions rares sous l'angle du repérage de traductions. Nous argumentons qu'elles sont plus difficiles à identifier que les traductions plus fréquentes. Nous décrivons une approche originale qui permet de mieux les identifier en tirant profit de l'alignement au niveau des mots de paires de phrases qui ne sont pas alignées. Nous montrons que cette approche permet d'améliorer l'identification de ces traductions rares.

Cet article présente MEltfr, un étiqueteur morpho-syntaxique automatique du français. Il repose sur un modèle probabiliste séquentiel qui bénéficie d'informations issues d'un lexique exogène, à savoir le Lefff. Evalué sur le FTB, MEltfr atteint un taux de précision de 97.75% (91.36% sur les mots inconnus) sur un jeu de 29 étiquettes. Ceci correspond à une diminution du taux d'erreur de 18% (36.1% sur les mots inconnus) par rapport au même modèle sans couplage avec le Lefff. Nous étudions plus en détail la contribution de cette ressource, au travers de deux séries d'expériences. Celles-ci font apparaître en particulier que la contribution des traits issus du Lefff est de permettre une meilleure couverture, ainsi qu'une modélisation plus fine du contexte droit des mots.

La définition de mesures sémantiques au niveau lexical a fait l'objet de nombreux travaux depuis plusieurs années. Dans cet article, nous nous focalisons plus spécifiquement sur les mesures de nature distributionnelle. Bien que différentes évaluations ont été réalisées les concernant, il reste difficile à établir si une mesure donnant de bons résultats dans un cadre d'évaluation peut être appliquée plus largement avec le même succès. Dans le travail présenté, nous commençons par sélectionner une mesure de similarité sur la base d'un test de type TOEFL étendu. Nous l'appliquons ensuite au problème de l'extraction de synonymes à partir de corpus en comparant nos

résultats avec ceux de (Curran & Moens, 2002). Enfin, nous testons l'intérêt pour cette tâche d'extraction de synonymes d'une méthode d'amélioration de la qualité des données distributionnelles proposée dans (Zhitomirsky-Geffet & Dagan, 2009).

Cet article décrit l'élaboration de la deuxième édition du dictionnaire de co-occurrences du logiciel d'aide à la rédaction Antidote. Cette nouvelle mouture est le résultat d'une refonte complète du processus d'extraction, ayant principalement pour but l'extraction de co-occurrences de plus de deux unités lexicales. La principale contribution de cet article est la description d'une technique originale pour l'extraction de co-occurrences de plus de deux mots conservant une structure syntaxique complète.

La récente éclosion du Web2.0 engendre un accroissement considérable de volumes textuels et intensifie ainsi l'importance d'une réflexion sur l'exploitation des connaissances à partir de grandes collections de documents. Dans cet article, nous présentons une approche de rechercher d'information qui s'inspire des certaines recherches issues de la psychologie cognitive pour la fouille de larges collections de documents. Nous utilisons un document comme requête permettant de récupérer des informations à partir d'une collection représentée dans un espace sémantique. Nous définissons les notions d'identité sémantique et de pollution sémantique dans un espace de documents. Nous illustrons notre approche par la description d'un système appelé BRAT (Blogosphere Random Analysis using Texts) basé sur les notions préalablement introduites d'identité et de pollution sématique appliquées à une tâche d'identification des actualités dans la blogosphère mondiale lors du concours TREC'09. Les premiers résultats produits sont tout à fait encourageant et indiquent les pistes des recherches à mettre en oeuvre afin d'améliorer les performances de BRAT.

Cet article propose une méthode pour calculer les dépendances syntaxiques d'un énoncé à partir du

processus d'analyse en constituants. L'objectif est d'obtenir des dépendances complètes c'est-à-dire contenant toutes les informations nécessaires à la construction de la sémantique. Pour l'analyse en constituants, on utilise le formalisme des grammaires d'interaction : celui-ci place au coeur de la composition syntaxique un mécanisme de saturation de polarités qui peut s'interpréter comme la réalisation d'une relation de dépendance. Formellement, on utilise la notion de motifs de graphes au sens de la réécriture de graphes pour décrire les conditions nécessaires à la création d'une dépendance.

Cet article présente une analyse statistique sur des données de syntaxe qui a pour but d'aider à mieux cerner le phénomène d'alternance de position de l'adjectif épithète par rapport au nom en français. Nous montrons comment nous avons utilisé les corpus dont nous disposons (French Treebank et le corpus de l'Est-Républicain) ainsi que les ressources issues du traitement automatique des langues, pour mener à bien notre étude. La modélisation à partir de 13 variables relevant principalement des propriétés du syntagme adjectival, de celles de l'item adjectival, ainsi que de contraintes basées sur la fréquence, permet de prédire à plus de 93% la position de l'adjectif. Nous insistons sur l'importance de contraintes relevant de l'usage pour le choix de la position de l'adjectif, notamment à travers la fréquence d'occurrence de l'adjectif, et la fréquence de contextes dans lesquels il apparaît.

Cet article présente un modèle de la complexité syntaxique. Il réunit un ensemble d'indices de complexité et les représente à l'aide d'un cadre formel homogène, offrant ainsi la possibilité d'une quantification automatique : le modèle proposé permet d'associer à chaque phrase un indice reflétant sa complexité.

Les structures de dépendances syntaxiques sont importantes et bien adaptées comme point de départ de diverses applications. Dans le cadre de l'analyseur TAG FRMG, nous présentons les

détails d'un processus de conversion de forêts partagées de dérivations en forêts partagées de dépendances. Des éléments d'information sont fournis sur un algorithme de désambiguisation sur ces forêts de dépendances.

Cet article présente une méthode hybride de normalisation des SMS, à mi-chemin entre correction orthographique et traduction automatique. La partie du système qui assure la normalisation utilise exclusivement des modèles entraînés sur corpus. Evalué en français par validation croisée, le système obtient un taux d'erreur au mot de 9.3% et un score BLEU de 0.83.

Dans cet article, nous introduisons une méthode à base de règles permettant d'extraire automatiquement de l'historique des éditions de l'encyclopédie collaborative Wikipédia des corrections orthographiques. Cette méthode nous a permis de construire un corpus d'erreurs composé de 72 483 erreurs lexicales (non-word errors) et 74 100 erreurs grammaticales (real-word errors). Il n'existe pas, à notre connaissance, de plus gros corpus d'erreurs écologiques librement disponible. En outre, les techniques mises en oeuvre peuvent être facilement transposées à de nombreuses autres langues. La collecte de ce corpus ouvre de nouvelles perspectives pour l'étude des erreurs fréquentes ainsi que l'apprentissage et l'évaluation des correcteurs orthographiques automatiques. Plusieurs expériences illustrant son intérêt sont proposées.

L'étiquetage sémantique consiste à associer un ensemble de propriétés à une séquence de mots contenue dans un texte. Bien que proche de la tâche d'étiquetage par entités nommées, qui revient à attribuer une classe de sens à un mot, la tâche d'étiquetage ou d'annotation sémantique cherche à établir la relation entre l'entité dans son texte et sa représentation ontologique. Nous présentons un étiqueteur sémantique qui s'appuie sur un étiqueteur d'entités nommées pour mettre en relation un mot ou un groupe de mots avec sa représentation ontologique. Son originalité est d'utiliser une ontologie intermédiaire de nature statistique pour établir ce lien.

Nous présentons le travail en cours effectué dans le cadre d'un projet d'extraction de paraphrases à partir de textes parallèles bilingues. Nous identifions des paraphrases sémantiques et lexico-syntaxiques, qui mettent en jeu des opérations relativement complexes sur les structures sémantiques et syntaxiques de phrases, et les décrivons au moyen de règles de paraphrasage de type Sens-Texte, utilisables dans diverses applications de TALN.

Cet article présente les premiers résultats d'une campagne d'annotation de corpus à grande échelle réalisée dans le cadre du projet ANNODIS. Ces résultats concernent la partie descendante du dispositif d'annotation, et plus spécifiquement les structures énumératives. Nous nous intéressons à la structuration énumérative en tant que stratégie de base de mise en texte, apparaissant à différents niveaux de granularité, associée à différentes fonctions discursives, et signalée par des indices divers. Avant l'annotation manuelle, une étape de pré-traitement a permis d'obtenir le marquage systématique de traits associés à la signalisation de l'organisation du discours. Nous décrivons cette étape de marquage automatique, ainsi que la procédure d'annotation. Nous proposons ensuite une première typologie des structures énumératives basée sur la description quantitative des données annotées manuellement, prenant en compte la couverture textuelle, la composition et les types d'indices.

Dans cet article, nous traitons de l'identification automatique des participants actants et circonstants de lexies prédicatives verbales tirées d'un corpus spécialisé en langue française. Les actants contribuent à la réalisation du sens de la lexie alors que les circonstants sont optionnels : ils ajoutent une information supplémentaire qui ne fait pas partie intégrante du sémantisme de la lexie. Nous proposons une classification de ces participants par apprentissage machine basée sur un corpus de lexies verbales du domaine de l'informatique, lexies qui ont été annotées manuellement avec des rôles sémantiques. Nous présentons des features qui nous permettent d'identifier les

participants et de distinguer les actants des circonstants.

Le but de ces travaux est d'extraire un lexique en analysant les relations entre des syntagmes nominaux et des syntagmes verbaux dans les textes de notre corpus, essentiellement des récits de voyage. L'hypothèse que nous émettons est de pouvoir établir une catégorisation des syntagmes nominaux associés à des Entités Nommées de type lieu à l'aide de l'analyse des relations verbales. En effet, nous disposons d'une chaine de traitement automatique qui extrait, interprète et valide des Entités Nommées de type lieu dans des documents textuels. Ce travail est complété par l'analyse des relations verbales associées à ces EN, candidates à l'enrichissement d'une ontologie.

Dans cet article nous proposons une nouvelle méthode pour l'identification du genre vidéo qui repose sur une analyse de leur contenu linguistique. Cette approche consiste en l'analyse des mots apparaissant dans les transcriptions des pistes audio des vidéos, obtenues à l'aide d'un système de reconnaissance automatique de la parole. Les expériences sont réalisées sur un corpus composé de dessins animés, de films, de journaux télévisés, de publicités, de documentaires, d'émissions de sport et de clips de musique. L'approche proposée permet d'obtenir un taux de bonne classification de 74% sur cette tâche. En combinant cette approche avec des méthodes reposant sur des paramètres acoustiques bas-niveau, nous obtenons un taux de bonne classification de 95%.

Les transcriptions automatiques de parole constituent une ressource importante, mais souvent bruitée, pour décrire des documents multimédia contenant de la parole (e.g. journaux télévisés). En vue d'améliorer la recherche documentaire, une étape d'extraction d'information à caractère sémantique, précédant l'indexation, permet de faire face au problème des transcriptions imparfaites. Parmis ces contenus informatifs, on compte les entités nommées (e.g. noms de personnes) dont l'extraction est l'objet de ce travail. Les méthodes traditionnelles de reconnaissance basées sur une définition manuelle de grammaires formelles donnent de bons résultats sur du texte ou des

transcriptions propres manuellement produites, mais leurs performances se trouvent fortement affectées lorsqu'elles sont appliquées sur des transcriptions automatiques. Nous présentons, ici, trois méthodes pour la reconnaissance d'entités nommées basées sur des algorithmes d'apprentissage automatique : les champs conditionnels aléatoires, les machines à de support, et les transducteurs à états finis. Nous présentons également une méthode pour rendre consistantes les données d'entrainement lorsqu'elles sont annotées suivant des conventions légèrement différentes. Les résultats montrent que les systèmes d'étiquetage obtenus sont parmi les plus robustes sur les données d'évaluation de la campagne ESTER 2 dans les conditions où la transcription automatique est particulièrement bruitée.

Les disfluences inhérents de toute parole spontanée sont un vrai défi pour les systèmes de compréhension de la parole. Ainsi, nous proposons dans cet article, une méthode originale pour le traitement des disfluences (plus précisément, les auto-corrections, les répétitions, les hésitations et les amorces) dans le cadre de la compréhension automatique de l'oral arabe spontané. Notre méthode est basée sur une analyse à la fois robuste et partielle, des énoncés oraux arabes. L'idée consiste à combiner une technique de reconnaissance de patrons avec une analyse sémantique superficielle par segments conceptuels. Cette méthode a été testée à travers le module de compréhension du système SARF, un serveur vocal interactif offrant des renseignements sur le transport ferroviaire tunisien (Bahou et al., 2008). Les résultats d'évaluation de ce module montrent que la méthode proposée est très prometteuse. En effet, les mesures de rappel, de précision et de F-Measure sont respectivement de 79.23%, 74.09% et 76.57%.

Les méthodes de segmentation thématique exploitant une mesure de la cohésion lexicale peuvent être appliquées telles quelles à des transcriptions automatiques de programmes télévisuels. Cependant, elles sont moins efficaces dans ce contexte, ne prenant en compte ni les particularités des émissions TV, ni celles des transcriptions. Nous étudions ici l'apport de relations sémantiques

pour rendre les techniques de segmentation thématique plus robustes. Nous proposons une méthode pour exploiter ces relations dans une mesure de la cohésion lexicale et montrons qu'elles permettent d'augmenter la F1-mesure de +1.97 et +11.83 sur deux corpus composés respectivement de 40h de journaux télévisés et de 40h d'émissions de reportage. Ces améliorations démontrent que les relations sémantiques peuvent rendre les méthodes de segmentation moins sensibles aux erreurs de transcription et au manque de répétitions constaté dans certaines émissions télévisées.

Cette étude a pour but de contribuer à la définition des objectifs de la segmentation thématique (ST), en incitant à prendre en considération le paramètre du type de textes dans cette tâche. Notre hypothèse est que, si la ST est certes pertinente pour traiter certains textes dont l'organisation est bien thématique, elle n'est pas adaptée à la prise en compte d'autres modes d'organisation (temporelle, rhétorique), et ne peut pas être appliquée sans précaution à des textes tout-venants. En comparant les performances d'un système de ST sur deux corpus, à organisation thématique "forte" et "faible", nous montrons que cette tâche est effectivement sensible à la nature des textes.

Dans le domaine de l'Extraction d'Information, une place importante est faite à l'extraction d'événements dans des dépêches d'actualité, particulièrement justifiée dans le contexte d'applications de veille. Or il est fréquent qu'une dépêche d'actualité évoque plusieurs événements de même nature pour les comparer. Nous proposons dans cet article d'étudier des méthodes pour segmenter les textes en séparant les événements, dans le but de faciliter le rattachement des informations pertinentes à l'événement principal. L'idée est d'utiliser des modèles d'apprentissage statistique exploitant les marqueurs temporels présents dans les textes pour faire cette segmentation. Nous présentons plus précisément deux modèles (HMM et CRF) entraînés pour cette tâche et, en faisant une évaluation de ces modèles sur un corpus de dépêches traitant d'événements sismiques, nous montrons que les méthodes proposées permettent d'obtenir des

résultats au moins aussi bons que ceux d'une approche ad hoc, avec une approche beaucoup plus générique.

Nous étudions différentes méthodes d'évaluation de résumé de documents basées sur le contenu. Nous nous intéressons en particulier à la corrélation entre les mesures d'évaluation avec et sans référence humaine. Nous avons développé FRESA, un nouveau système d'évaluation fondé sur le contenu qui calcule les divergences entre les distributions de probabilité. Nous appliquons notre système de comparaison aux diverses mesures d'évaluation bien connues en résumé de texte telles que la Couverture, Responsiveness, Pyramids et Rouge en étudiant leurs associations dans les tâches du résumé multi-document générique (francais/anglais), focalisé (anglais) et résumé mono-document générique (français/espagnol).

La majorité des systèmes de résumés automatiques sont basés sur l'extraction de phrases, or on les compare le plus souvent avec des résumés rédigés manuellement par abstraction. Nous avons mené une expérience dans le but d'établir une limite supérieure aux performances auxquelles nous pouvons nous attendre avec une approche par extraction. Cinq résumeurs humains ont composé 88 résumés de moins de 100 mots, en extrayant uniquement des phrases présentes intégralement dans les documents d'entrée. Les résumés ont été notés sur la base de leur contenu, de leur niveau linguistique et de leur qualité globale par les évaluateurs de NIST dans le cadre de la compétition TAC 2009. Ces résumés ont obtenus de meilleurs scores que l'ensemble des 52 systèmes automatiques participant à la compétition, mais de nettement moins bons que ceux obtenus par les résumeurs humains pouvant formuler les phrases de leur choix dans le résumé. Ce grand écart montre l'insuffisance des méthodes par extraction pure.

Cet article présente l'étude d'un corpus de réponses formulées par des humains à des questions factuelles. Des observations qualitatives et quantitatives sur la reprise d'éléments de la question

dans les réponses sont exposées. La notion d'information-réponse est introduite et une étude de la présence de cet élément dans le corpus est proposée. Enfin, les formulations des réponses sont étudiées.

Cet article présente une méthode non-supervisée pour extraire des paires de phrases parallèles à partir d'un corpus comparable. Un système de traduction automatique est utilisé pour exploiter le corpus comparable et détecter les paires de phrases parallèles. Un processus itératif est exécuté non seulement pour augmenter le nombre de paires de phrases parallèles extraites, mais aussi pour améliorer la qualité globale du système de traduction. Une comparaison avec une méthode semi-supervisée est présentée également. Les expériences montrent que la méthode non-supervisée peut être réellement appliquée dans le cas où on manque de données parallèles. Bien que les expériences préliminaires soient menées sur la traduction français-anglais, cette méthode non-supervisée est également appliquée avec succès à un couple de langues peu doté : vietnamien-français.

De nombreux travaux en Traduction Automatique Statistique (TAS) pour des langues d'entrée morphologiquement riches montrent que la ségmentation morphologique et la normalisation orthographique améliorent la qualité des traductions en diminuant la sparsité des données. Dans cet article, nous étudions l'impact de ce pré-traitement pour la TAS vers une langue de sortie riche morphologiquement, comme l'Arabe. Nous explorons l'espace des schémas de segmentation et des options de normalisation possibles. Nous évaluons seulement la sortie sous une forme désegmentée et enrichie orthographiquement. Nos résultats montrent d'une part que le meilleur schéma pour la ségmentation est celui de la Penn Arabic Treebank. D'autre part, la meilleure procédure de pré-traitement consiste à entraîner le système sur des données normalisées orthographiquement, puis à enrichir et désegmenter les traductions en sortie.

Distinguer les constructions verbe-sujet (VS) des propositions principales ("matrice") et subordonnées ("non-matrice") améliore notre nouveau modèle de ré-ordonnancement pour l'alignement des mots en Traduction Automatique Statistique (TAS) arabe-anglais (Carpuat et al., 2010). D'une part, la majorité des constructions verbe-sujet (VS) dans les propositions principales doivent être réordonnancées en anglais, alors que l'ordre du verbe et du sujet est préservé dans la moitié des cas de constructions VS subordonnées. D'autre part, nous constatons que notre analyseur syntaxique parvient à mieux identifier les constructions VS des propositions principales. Ces observations nous amènent à limiter le ré-ordonnancement des constructions VS à celles des propositions principales lors de l'alignement des mots. Cette technique améliore substantiellement la performance d'un système de TAS conventionnel, et d'un système qui réordonnance toutes les constructions VS. L'amélioration des mesures BLEU et TER obtenue par simple ré-ordonnancement représente presque la moitié de l'amélioration obtenue lorsque le modèle d'alignement des mots est entraîné sur un corpus parallèle d'une taille cinq fois supérieure.

Historiquement deux types de traitement de la langue ont été étudiés: le traitement par le cerveau (approche psycholinguistique) et le traitement par la machine (approche TAL). Nous pensons qu'il y a place pour un troisième type: le traitement interactif de la langue (TIL), l'ordinateur assistant le cerveau. Ceci correspond à un besoin réel dans la mesure où les gens n'ont souvent que des connaissances partielles par rapport au problème à résoudre. Le but du TIL est de construire des ponts entre ces connaissances momentanées d'un utilisateur et la solution recherchée. À l'aide de quelques exemples, nous essayons de montrer que ceci est non seulement faisable et souhaitable, mais également d'un coût très raisonnable.

Cet article décrit des modifications du dictionnaire de valence des verbes du français DICOVALENCE qui visent à le rendre neutre par rapport aux modèles syntaxiques, à expliciter certaines informations sur le cadre de sous-catégorisation et à le rendre ainsi directement utilisable

en TAL. Les informations explicitées sont les suivantes : (a) les fonctions syntaxiques des arguments verbaux, (b) les restrictions de sélection portant sur ces arguments et (c) leurs réalisations syntagmatiques possibles. Les restrictions sont exprimées à l'aide de traits sémantiques. L'article décrit aussi le calcul de ces traits sémantiques à partir des paradigmes des pronoms (et d'éléments similaires) associés aux arguments. On obtient un format indépendant du modèle syntaxique, dont l'interprétation est transparente.

Notre recherche démontre que l'utilisation du contenu d'un texte à traduire permet de mieux cibler dans une banque de terminologie les équivalents terminologiques pertinents à ce texte. Une banque de terminologie a comme particularité qu'elle catégorise ses entrées (fiches) en leur assignant un ou des domaines provenant d'une liste de domaines préétablie. La stratégie ici présentée repose sur l'utilisation de cette information sur les domaines. Un algorithme a été développé pour l'assignation automatique d'un profil de domaines à un texte. Celui-ci est combiné à un algorithme d'appariement entre les domaines d'un terme présent dans la banque de terminologie et le profil de domaines du texte. Pour notre expérimentation, des résumés bilingues (français et anglais) provenant de huit revues scientifiques nous fournissent un ensemble de 1130 paires d'équivalents terminologiques et le Grand Dictionnaire Terminologique (Office Québécois de la Langue Française) nous sert de ressource terminologique. Sur notre ensemble, nous démontrons une réduction de 75% du rang moyen de l'équivalent correct en comparaison avec un choix au hasard.

Nous définissons le ?-calcul, un calcul de réécriture de graphes, que nous proposons d'utiliser pour étudier les liens entre différentes représentations linguistiques. Nous montrons comment transformer une analyse syntaxique en une représentation sémantique par la composition de deux jeux de règles de beta-calcul. Le premier souligne l'importance de certaines informations syntaxiques pour le calcul de la sémantique et explicite le lien entre syntaxe et sémantique sous-spécifiée. Le second décompose la recherche de modèles pour les représentations

sémantiques sous-spécifiées.

L'objectif des travaux présentés dans cet article est l'évaluation de la qualité d'annotations manuelles de relations de renommage de gènes dans des résumés scientifiques, annotations qui présentent la caractéristique d'être très dispersées. Pour cela, nous avons calculé et comparé les coefficients les plus communément utilisés, entre autres kappa (Cohen, 1960) et pi (Scott, 1955), et avons analysé dans quelle mesure ils sont adaptés à nos données. Nous avons également étudié les différentes pondérations applicables à ces coefficients permettant de calculer le kappa pondéré (Cohen, 1968) et l'alpha (Krippendorff, 1980, 2004). Nous avons ainsi étudié le biais induit par la grande prévalence d'une catégorie et défini un mode de calcul des distances entre catégories reposant sur les annotations réalisées.

Nous présentons dans cet article une étude empirique de l'application de l'approche de l'entropie maximale pour l'étiquetage syntaxique de textes vietnamiens. Le vietnamien est une langue qui possède des caractéristiques spéciales qui la distinguent largement des langues occidentales. Notremeilleur étiqueteur explore et inclut des connaissances utiles qui, en terme de performance pour l'étiquetage de textes vietnamiens, fournit un taux de précision globale de 93.40% et de 80.69% pour les mots inconnus sur un ensemble de test du corpus arboré vietnamien. Notre étiqueteur est nettement supérieur à celui qui est en train d'être utilisé pour développer le corpus arboré vietnamien, et à l'heure actuelle c'est le meilleur résultat obtenu pour l'étiquetage de textes vietnamiens.

Le projet décrit vise à soutenir les efforts de constitution de ressources lexicales utiles à l'indexation automatique. Un type de vocabulaire utile à l'indexation est défini, le vocabulaire savant de base, qui peut s'articuler avec le vocabulaire spécialisé pour constituer des entrées d'index structurées. On présente les résultats d' une expérimentation d' extraction (semi-)automatique des mots du

vocabulaire savant de base à partir d'un corpus ciblé, constitué de résumés d'articles scientifiques en franc?ais et en anglais. La tâche d'extraction a réussi à doubler une liste originale constituée manuellement pour le français. La comparaison est établie avec une expérimentation similaire effectuée pour l'anglais sur un corpus plus grand et contenant des résumés d'articles non seulement en sciences pures mais aussi en sciences humaines et sociales.

Bien que les approches fondées sur la théorie de l'information sont prédominantes dans le domaine de l'analyse morphologique non supervisée, depuis quelques années, d'autres approches ont gagné en popularité, dont celles basées sur l'analogie formelle. Cette dernière reste tout de même marginale due notamment à son coût de calcul élevé. Dans cet article, nous proposons un algorithme basé sur l'analogie formelle capable de traiter les lexiques volumineux. Nous introduisons pour cela le concept de règle de cofacteur qui permet de généraliser l'information capturée par une analogie tout en contrôlant les temps de traitement. Nous comparons notre système à 2 systèmes : Morfessor (Creutz & Lagus, 2005), un système de référence dans de nombreux travaux sur l'analyse morphologique et le système analogique décrit par Langlais (2009). Nous en montrons la supériorité pour 3 des 5 langues étudiées ici : le finnois, le turc, et l'allemand.

Dans le domaine biomédical, beaucoup de termes sont des composés savants (composés de plusieurs racines gréco-latines). L'étude de leur morphologie est importante pour de nombreuses applications puisqu'elle permet de structurer ces termes, de les rechercher efficacement, de les traduire... Dans cet article, nous proposons de suivre une démarche originale mais fructueuse pour mener cette analyse morphologique sur des termes simples en français, en nous appuyant sur une langue pivot, le japonais, et plus précisément sur les termes écrits en kanjis. Pour cela nous avons développé un algorithme d'alignement de termes spécialement adapté à cette tâche. C'est cet alignement d'un terme français avec sa traduction en kanjis qui fournit en même temps une décomposition en morphe et leur étiquetage par les kanjis correspondants. Évalué sur un jeu de

données conséquent, notre approche obtient une précision supérieure à 70% et montrent son bien fondé en comparaison avec les techniques existantes. Nous illustrons également l'intérêt de notre démarche au travers de deux applications directes de ces alignements : la traduction de termes inconnus et la découverte de relations entre morphes pour la tructuration terminologique.

Nous présentons PerLex, un lexique morphologique du persan à large couverture et librement disponible, accompagné d'une chaîne de traitements de surface pour cette langue. Nous décrivons quelques caractéristiques de la morphologie du persan, et la façon dont nous l'avons représentée dans le formalisme lexical Alexina, sur lequel repose PerLex. Nous insistons sur la méthodologie que nous avons employée pour construire les entrées lexicales à partir de diverses sources, ainsi que sur les problèmes liés à la normalisation typographique. Le lexique obtenu a une couverture satisfaisante sur un corpus de référence, et devrait donc constituer un bon point de départ pour le développement d'un lexique syntaxique du persan.

Cet article concerne la détermination de la similarité entre des textes courts (phrases, paragraphes, ...). Ce problème est souvent abordé dans la littérature à l'aide de méthodes supervisées ou de ressources externes comme le thesaurus Wordnet ou le British National Corpus. Les méthodes que nous proposons sont non supervisées et n'utilisent pas de connaissances à priori. La première méthode que nous présentons est basée sur le modèle vectoriel de Salton auquel nous avons apporté des modifications pour prendre en compte le contexte, le sens et la relation entre les mots des textes. Dans un deuxième temps, nous testons les mesures de Dice et de ressemblance pour résoudre ce problème ainsi que l'utilisation de la racinisation. Enfin, ces différentes méthodes sont évaluées et comparées aux résultats obtenus dans la littérature.

Cet article présente une méthodologie d'utilisation du Trésor de la Langue Française informatisée (TLFi) pour l'indexation et la recherche des images fondée sur l'annotation textuelle. Nous utilisons

les définitions du TLFi pour la création automatique et l'enrichissement d'un thésaurus à partir des mots-clés de la requête de recherche et des mots-clés attribués à l'image lors de l'indexation. Plus précisement il s'agit d'associer, de façon automatisé, à chaque mot-clé de l'image une liste des mots extraits de ses définitions TLFi pour un domaine donné, en construisant ainsi un arbre hiérarchique. L'approche proposée permet une catégorisation très précise des images, selon les domaines, une indexation de grandes quantités d'images et une recherche rapide.

Ce papier présente une méthode de recherche des phrases évaluatives dans les articles de presse économique et financière à partir de marques et d'indices stéréotypés, propres au style journalistique, apparaissant de manière concomitante à l'expression d'évaluation(s) dans les phrases. Ces marques et indices ont été dégagés par le biais d'une annotation manuelle. Ils ont ensuite été implémentés, en vue d'une phase-test d'annotation automatique, sous forme de grammaires DCG/GULP permettant, par filtrage, de matcher les phrases les contenant. Les résultats de notre première tentative d'annotation automatique sont présentés dans cet article. Enfin les perspectives offertes par cette méthode relativement peu coûteuse en ressources (à base d'indices non intrinsèquement évaluatifs) font l'objet d'une discussion.

La complexité linguistique regroupe différents phénomènes dont il s'agit de modéliser le rapport. Le travail en cours que je décris ici propose une réflexion sur les approches linguistiques et techniques de cette notion et la mise en application d'un balayage des textes qui s'efforce de contribuer à leur enrichissement. Ce traitement en surface effectué suivant une liste de critères qui représentent parfois des approximations de logiques plus élaborées tente de fournir une image raisonnable de la complexité.

Dans cet article, nous cherchons à identifier la nature de l'ambiguïté des requêtes utilisateurs issues d'un moteur de recherche dédié à l'actualité, 2424actu.fr, en utilisant une tâche de catégorisation.

Dans un premier temps, nous verrons les différentes formes de l'ambiguïté des requêtes déjà décrites dans les travaux de TAL. Nous confrontons la vision lexicographique de l'ambiguïté à celle décrite par les techniques de classification appliquées à la recherche d'information. Dans un deuxième temps, nous appliquons une méthode de catégorisation thématique afin d'explorer l'ambiguïté des requêtes, celle-ci nous permet de conduire une analyse sémantique de ces requêtes, en intégrant la dimension temporelle propre au contexte des news. Nous proposons une typologie des phénomènes d'ambiguïté basée sur notre analyse sémantique. Enfin, nous comparons l'exploration par catégorisation à une ressource comme Wikipédia, montrant concrètement les divergences des deux approches.

Plusieurs utilisateurs ont souvent besoin d'informations pédagogiques pour les intégrer dans leurs ressources pédagogiques, ou pour les utiliser dans un processus d'apprentissage. Une indexation de ces informations s'avère donc utile en vue d'une extraction des informations pédagogiques pertinentes en réponse à une requête utilisateur. La plupart des systèmes d'extraction d'informations pédagogiques existants proposent une indexation basée sur une annotation manuelle ou semi-automatique des informations pédagogiques, tâche qui n'est pas préférée par les utilisateurs. Dans cet article, nous proposons une approche d'indexation d'objets pédagogiques (Définition, Exemple, Exercice, etc.) basée sur une annotation sémantique par Exploration Contextuelle des documents. L'index généré servira à une extraction des objets pertinents répondant à une requête utilisateur sémantique. Nous procédons, ensuite, à un classement des objets extraits selon leur pertinence en utilisant l'algorithme Rocchio. Notre objectif est de mettre en valeur une indexation à partir de contextes sémantiques et non pas à partir de seuls termes linguistiques.

Cet article présente des applications d'outils et méthodes du traitement automatique des langues (TAL) à la maîtrise du risque industriel grâce à l'analyse de données textuelles issues de

volumineuses bases de retour d'expérience (REX). Il explicite d'abord le domaine de la gestion de la sûreté, ses aspects politiques et sociaux ainsi que l'activité des experts en sûreté et les besoins qu'ils expriment. Dans un deuxième temps il présente une série de techniques, comme la classification automatique de documents, le repérage de subjectivité, et le clustering, adaptées aux données REX visant à répondre à ces besoins présents et à venir, sous forme d'outils, en support à l'activité des experts.

Nous proposons une méthodologie pour la construction de règles de déduction de relations de discours, destinées à être intégrées dans une algèbre de ces relations. La construction de ces règles a comme principal objectif de pouvoir calculer la fermeture discursive d'une structure de discours, c'est-à-dire de déduire toutes les relations que la structure contient implicitement. Calculer la fermeture des structures discursives peut permettre d'améliorer leur comparaison, notamment dans le cadre de l'évaluation de systèmes d'analyse automatique du discours. Nous présentons la méthodologie adoptée, que nous illustrons par l'étude d'une règle de déduction.

Les corpus comparables monolingues, alignés non pas au niveau des documents mais au niveau d'unités textuelles plus fines (paragraphe, phrases, etc.), sont utilisés dans diverses applications de traitement automatique des langues comme par exemple en détection de plagiat. Mais ces types de corpus ne sont pratiquement pas disponibles et les chercheurs sont donc obligés de les construire et de les annoter manuellement, ce qui est un travail très fastidieux et coûteux en temps. Dans cet article, nous présentons une méthode, composée de deux étapes, qui permet de réduire ce travail d'annotation de segments de texte. Cette méthode est évaluée lors de l'alignement de paragraphes provenant de dépêches en langue anglaise issues de diverses sources. Les résultats obtenus montrent un apport considérable de la méthode en terme de réduction de temps d'annotation. Nous présentons aussi des premiers résultats obtenus à l'aide de simples traitements automatiques (recouvrement de mots, de racines, mesure cosinus) pour tenter de diminuer encore la charge de

travail humaine.

Détection d'émotion, fouille d'opinion et analyse des sentiments sont généralement évalués par comparaison des réponses du système concerné par rapport à celles contenues dans un corpus de référence. Les questions posées dans cet article concernent à la fois la définition de la référence et la fiabilité des métriques les plus fréquemment utilisées pour cette comparaison. Les expérimentations menées pour évaluer le système de détection d'émotions EmoLogus servent de base de réflexion pour ces deux problèmes. L'analyse des résultats d'EmoLogus et la comparaison entre les différentes métriques remettent en cause le choix du vote majoritaire comme référence. Par ailleurs elles montrent également la nécessité de recourir à des outils statistiques plus évolués que ceux généralement utilisés pour obtenir des évaluations fiables de systèmes qui travaillent sur des données intrinsèquement subjectives et incertaines.

Dans cet article, nous décrivons une nouvelle approche pour la création de résumés extractifs ? tâche qui consiste à créer automatiquement un résumé pour un document en sélectionnant un sous-ensemble de ses phrases ? qui exploite des informations collocationnelles spécifiques à un domaine, acquises préalablement à partir d'un corpus de développement. Un extracteur de collocations fondé sur l'analyse syntaxique est utilisé afin d'inférer un modèle de contenu qui est ensuite appliqué au document à résumer. Cette approche a été utilisée pour la création des versions simples pour les articles de Wikipedia en anglais, dans le cadre d'un projet visant la création automatique d'articles simplifiées, similaires aux articles recensées dans Simple English Wikipedia. Une évaluation du système développé reste encore à faire. Toutefois, les résultats préalables obtenus pour les articles sur des villes montrent le potentiel de cette approche guidée par collocations pour la sélection des phrases pertinentes.

Cette étude envisage l'emploi des unités polylexicales (UPs) comme prédicteurs dans une formule

de lisibilité pour le français langue étrangère. À l'aide d'un extracteur d'UPs combinant une approche statistique à un filtre linguistique, nous définissons six variables qui prennent en compte la densité et la probabilité des UPs nominales, mais aussi leur structure interne. Nos expérimentations concluent à un faible pouvoir prédictif de ces six variables et révèlent qu'une simple approche basée sur la probabilité moyenne des n-grammes des textes est plus efficace.

La détection et le typage des entités nommées sont des tâches pour lesquelles ont été développés à la fois des systèmes symboliques et probabilistes. Nous présentons les résultats d'une expérience visant à faire interagir le système à base de règles NP, développé sur des corpus provenant de l'AFP, intégrant la base d'entités Aleda et qui a une bonne précision, et le système LIANE, entraîné sur des transcriptions de l'oral provenant du corpus ESTER et qui a un bon rappel. Nous montrons qu'on peut adapter à un nouveau type de corpus, de manière non supervisée, un système probabiliste tel que LIANE grâce à des corpus volumineux annotés automatiquement par NP. Cette adaptation ne nécessite aucune annotation manuelle supplémentaire et illustre la complémentarité des méthodes numériques et symboliques pour la résolution de tâches linguistiques.

La constitution de ressources linguistiques est une tâche longue et coûteuse. C'est notamment le cas pour les ressources morphologiques. Ces ressources décrivent de façon approfondie et explicite l'organisation morphologique du lexique complétée d'informations sémantiques exploitables dans le domaine du TAL. Le travail que nous présentons dans cet article s'inscrit dans cette perspective et, plus particulièrement, dans l'optique d'affiner une ressource existante en s'appuyant sur des informations sémantiques obtenues automatiquement. Notre objectif est de caractériser sémantiquement des familles morpho-phonologiques (des mots partageant une même racine et une continuité de sens). Pour ce faire, nous avons utilisé des informations extraites du TLFi annoté morpho-syntaxiquement. Les premiers résultats de ce travail seront analysés et discutés.

Nous présentons dans cet article un générateur automatique de questions pour le français. Le système de génération procède par transformation de phrases déclaratives en interrogatives et se base sur une analyse syntaxique préalable de la phrase de base. Nous détaillons les différents types de questions générées. Nous présentons également une évaluation de l'outil, qui démontre que 41 % des questions générées par le système sont parfaitement bien formées.

Les systèmes de questions réponses recherchent la réponse à une question posée en langue naturelle dans un ensemble de documents. Les collectionsWeb diffèrent des articles de journaux de par leurs structures et leur style. Pour tenir compte de ces spécificités nous avons développé un système fondé sur une approche robuste de validation où des réponses candidates sont extraites à partir de courts passages textuels puis ordonnées par apprentissage. Les résultats montrent une amélioration du MRR (Mean Reciprocal Rank) de 48% par rapport à la baseline.

Le domaine de l'extraction d'information s'est récemment développé en limitant les contraintes sur la définition des informations à extraire, ouvrant la voie à des applications de veille plus ouvertes. Dans ce contexte de l'extraction d'information non supervisée, nous nous intéressons à l'identification et la caractérisation de nouvelles relations entre des types d'entités fixés. Un des défis de cette tâche est de faire face à la masse importante de candidats pour ces relations lorsque l'on considère des corpus de grande taille. Nous présentons dans cet article une approche pour le filtrage des relations combinant méthode heuristique et méthode par apprentissage. Nous évaluons ce filtrage de manière intrinsèque et par son impact sur un regroupement sémantique des relations.

Cet article décrit une étude sur l'annotation automatique des noms d'événements dans les textes en français. Plusieurs lexiques existants sont utilisés, ainsi que des règles syntaxiques d'extraction, et un lexique composé de façon automatique, permettant de fournir une valeur sur le niveau d'ambiguïté du mot en tant qu'événement. Cette nouvelle information permettrait d'aider à la

désambiguïsation des noms d'événements en contexte.

Les approches statistiques les plus performantes actuellement pour la compréhension automatique du langage naturel nécessitent une annotation segmentale des données d'entraînement. Nous étudions dans cet article une alternative permettant d'obtenir de façon non-supervisée un alignement segmental d'unités conceptuelles sur les mots. L'impact de l'alignement automatique sur les performances du système de compréhension est évalué sur une tâche de dialogue oral.

La recherche de passages consiste à extraire uniquement des passages pertinents par rapport à une requête utilisateur plutôt qu'un ensemble de documents entiers. Cette récupération de passages est souvent handicapée par le manque d'informations complémentaires concernant le contexte de la recherche initiée par l'utilisateur. Des études montrent que l'ajout d'informations contextuelles par l'utilisateur peut améliorer les performances des systèmes de recherche de passages. Nous confirmons ces observations dans cet article, et nous introduisons également une méthode d'enrichissement de la requête à partir d'informations contextuelles issues de documents encyclopédiques. Nous menons des expérimentations en utilisant la collection et les méthodes d'évaluation proposées par la campagne INEX. Les résultats obtenus montrent que l'ajout d'informations contextuelles permet d'améliorer significativement les performances de notre système de recherche de passages. Nous observons également que notre approche automatique obtient les meilleurs résultats parmi les différentes approches que nous évaluons.

Après une brève analyse linguistique des adjectifs dénominaux en français, nous décrivons le processus automatique que nous avons mis en place à partir de lexiques et de corpus volumineux pour construire un lexique d'adjectifs dénominaux dérivés de manière régulière. Nous estimons à la fois la précision et la couverture du lexique dérivationnel obtenu. À terme, ce lexique librement disponible aura été validé manuellement et contiendra également les adjectifs dénominaux à base

supplétive.

Nous présentons une nouvelle version de PerLex, lexique morphologique du persan, une version corrigée et partiellement réannotée du corpus étiqueté BijanKhan (BijanKhan, 2004) et MEltfa, un nouvel étiqueteur morpho-syntaxique librement disponible pour le persan. Après avoir développé une première version de PerLex (Sagot & Walther, 2010), nous en proposons donc ici une version améliorée. Outre une validation manuelle partielle, PerLex 2 repose désormais sur un inventaire de catégories linguistiquement motivé. Nous avons également développé une nouvelle version du corpus BijanKhan : elle contient des corrections significatives de la tokenisation ainsi qu'un réétiquetage à l'aide des nouvelles catégories. Cette nouvelle version du corpus a enfin été utilisée pour l'entraînement de MEltfa, notre étiqueteur morpho-syntaxique pour le persan librement disponible, s'appuyant à la fois sur ce nouvel inventaire de catégories, sur PerLex 2 et sur le système d'étiquetage MElt (Denis & Sagot, 2009).

Cet article présente une méthode hybride d'identification de cognats français - roumain. Cette méthode exploite des corpus parallèles alignés au niveau propositionnel, lemmatisés et étiquetés (avec des propriétés morpho-syntaxiques). Notre méthode combine des techniques statistiques et des informations linguistiques pour améliorer les résultats obtenus. Nous évaluons le module d'identification de cognats et nous faisons une comparaison avec des méthodes statistiques pures, afin d'étudier l'impact des informations linguistiques utilisées sur la qualité des résultats obtenus. Nous montrons que l'utilisation des informations linguistiques augmente significativement la performance de la méthode.

Ce papier s'inscrit dans le cadre général de l'Apprentissage et de l'Enseignement des Langues Assistés par Ordinateur, et concerne plus particulièrement l'automatisation des exercices de dictée. Il présente une méthode de correction des copies d'apprenants qui se veut originale en deux points.

Premièrement, la méthode exploite la composition d'automates à états finis pour détecter et pour analyser les erreurs. Deuxièmement, elle repose sur une analyse morpho-syntaxique automatique de l'original de la dictée, ce qui facilite la production de diagnostics.

Dans cet article, nous présentons la définition et l'étude d'un corpus de dialogues entre un schizophrène et un interlocuteur ayant pour objectif la conduite et le maintien de l'échange. Nous avons identifié des discontinuités significatives chez les schizophrènes paranoïdes. Une représentation issue de la S-DRT (sa partie pragmatique) permet de rendre compte des ces usages non standards.

Cet article présente un corpus parallèle français-allemand de plus de 4 millions de mots issu de la numérisation d'un corpus alpin multi-lingue. Ce corpus est une précieuse ressource pour de nombreuses études de linguistique comparée et du patrimoine culturel ainsi que pour le développement d'un système statistique de traduction automatique dans un domaine spécifique. Nous avons annoté un échantillon de ce corpus parallèle et aligné les structures arborées au niveau des mots, des constituants et des phrases. Cet "alpine treebank" est le premier corpus arboré parallèle français-allemand de haute qualité (manuellement contrôlé), de libre accès et dans un domaine et un genre nouveau : le récit d'alpinisme.

Nous présentons différentes méthodes de ré-ordonnancement de phrases pour le résumé automatique fondé sur une classification des phrases à résumer en classes thématiques. Nous comparons ces méthodes à deux baselines : ordonnancement des phrases selon leur pertinence et ordonnancement selon la date et la position dans le document d'origine. Nous avons fait évaluer les résumés obtenus sur le corpus RPM2 par 4 annotateurs et présentons les résultats.

La langue arabe présente des spécificités qui la rendent plus ambigu?e que d'autres langues

naturelles. Sa morphologie, sa syntaxe ainsi que sa sémantique sont en corrélation et se complètent l'une l'autre. Dans le but de construire une grammaire qui soit adaptée à ces spécificités, nous avons conçu et développé une application d'aide à la création des règles syntaxiques licites suivant le formalisme d'arbres adjoints. Cette application est modulaire et enrichie par des astuces de contrôle de la création et aussi d'une interface conviviale pour assister l'utilisateur final dans la gestion des créations prévues.

Dans cet article, nous présentons FreDist, un logiciel libre pour la construction automatique de thésaurus distributionnels à partir de corpus de texte, ainsi qu'une évaluation des différents ressources ainsi produites. Suivant les travaux de (Lin, 1998) et (Curran, 2004), nous utilisons un corpus journalistique de grande taille et implémentons différentes options pour : le type de relation contexte lexical, la fonction de poids, et la fonction de mesure de similarité. Prenant l'EuroWord-Net français et le WOLF comme références, notre évaluation révèle, de manière originale, que c'est l'approche qui combine contextes linéaires (ici, de type bi-grammes) et contextes syntaxiques qui semble fournir le meilleur thésaurus. Enfin, nous espérons que notre logiciel, distribué avec nos meilleurs thésaurus pour le français, seront utiles à la communauté TAL.

Dans cet article, nous proposons de modéliser la tâche d'extraction de relations à partir de corpus textuels comme un problème de classification. Nous montrons que, dans ce cadre, des représentations fondées sur des informations linguistiques de surface sont suffisantes pour que des algorithmes d'apprentissage artificiel standards les exploitant rivalisent avec les meilleurs systèmes d'extraction de relations reposant sur des connaissances issues d'analyses profondes (analyses syntaxiques ou sémantiques). Nous montrons également qu'en prenant davantage en compte les spécificités de la tâche d'extraction à réaliser et des données disponibles, il est possible d'obtenir des méthodes encore plus efficaces tout en exploitant ces informations simples. La technique originale à base d'apprentissage « paresseux » et de modèles de langue que nous évaluons en

extraction d'interactions géniques sur les données du challenge LLL2005 dépasse les résultats de l'état de l'art.

Les travaux menés ces dernières années autour de l'itération en langue, tant par la communauté linguistique que par celle du TAL, ont mis au jour des phénomènes particuliers, non réductibles aux représentations temporelles classiques. En particulier, une itération ne saurait structurellement être réduite à une simple énumération de procès, et du point de vue de l'aspect, met en jeu simultanément deux visées aspectuelles indépendantes. Le formalisme TimeML, qui a vocation à annoter les informations temporelles portées par un texte, intègre déjà des éléments relatifs aux itérations, mais ne prend pas en compte ces dernières avancées. C'est ce que nous entreprenons de faire dans cet article, en proposant une extension à ce formalisme.

Cette recherche s'inscrit dans le champ de la fouille d'opinion et, plus particulièrement, dans celui de l'analyse de la polarité d'une phrase ou d'un syntagme. Dans ce cadre, la prise en compte du contexte linguistique dans lequel apparaissent les mots porteurs de valence est particulièrement importante. Nous proposons une méthodologie pour extraire automatiquement de corpus de textes de telles expressions linguistiques. Cette approche s'appuie sur un corpus de textes, ou d'extraits de textes, dont la valence est connue, sur un lexique de valence construit à partir de ce corpus au moyen d'une procédure automatique et sur un analyseur syntaxique. Une étude exploratoire, limitée à la seule relation syntaxique associant un adverbe à un adjectif, laisse entrevoir les potentialités de l'approche.

La multiplication des travaux sur corpus, en linguistique computationnelle et en TAL, conduit à la multiplication des campagnes d'annotation et des corpus multi-annotés, porteurs d'informations relatives à des phénomènes variés, envisagés par des annotateurs multiples, parfois automatiques. Pour mieux comprendre les phénomènes que ces campagnes prennent pour objets, ou pour

contrôler les données en vue de l'établissement d'un corpus de référence, il est nécessaire de disposer d'outils permettant d'explorer les annotations. Nous présentons une stratégie possible et son opérationalisation dans la plate-forme Glozz par le langage GlozzQL.

Dans cet article, nous traitons de l'attribution des rôles sémantiques aux actants de lexies verbales en corpus spécialisé en français. Nous proposons une classification de rôles sémantiques par apprentissage machine basée sur un corpus de lexies verbales annotées manuellement du domaine de l'informatique et d'Internet. Nous proposons également une méthode de partitionnement semi-supervisé pour prendre en compte l'annotation de nouvelles lexies ou de nouveaux rôles sémantiques et de les intégrés dans le système. Cette méthode de partitionnement permet de regrouper les instances d'actants selon les valeurs communes correspondantes aux traits de description des actants dans des groupes d'instances d'actants similaires. La classification de rôles sémantique a obtenu une F-mesure de 93% pour Patient, de 90% pour Agent, de 85% pour Destination et de 76% pour les autres rôles pris ensemble. Quand au partitionnement en regroupant les instances selon leur similarité donne une F-mesure de 88% pour Patient, de 81% pour Agent, de 58% pour Destination et de 46% pour les autres rôles.

Les travaux sur les mesures de similarité sémantique de nature distributionnelle ont abouti à un certain consensus quant à leurs performances et ont montré notamment que leurs résultats sont surtout intéressants pour des mots de forte fréquence et une similarité sémantique étendue, non restreinte aux seuls synonymes. Dans cet article, nous proposons une méthode d'amélioration d'une mesure de similarité classique permettant de rééquilibrer ses résultats pour les mots de plus faible fréquence. Cette méthode est fondée sur un mécanisme d'amorçage : un ensemble d'exemples et de contre-exemples de mots sémantiquement liés sont sélectionnés de façon non supervisée à partir des résultats de la mesure initiale et servent à l'entraînement d'un classifieur supervisé. Celui-ci est ensuite utilisé pour réordonner les voisins sémantiques initiaux. Nous évaluons l'intérêt

de ce ré-ordonnancement pour un large ensemble de noms anglais couvrant différents domaines fréquentiels.

Ce travail s'inscrit dans l'analyse automatique d'un corpus de récits de voyage. À cette fin, nous raffinons la sémantique de Montague pour rendre compte des phénomènes d'adaptation du sens des mots au contexte dans lequel ils apparaissent. Ici, nous modélisons les constructions de type 'le chemin descend pendant une demi-heure' où ledit chemin introduit un voyageur fictif qui le parcourt, en étendant des idées que le dernier auteur a développé avec Bassac et Mery. Cette introduction du voyageur utilise la montée de type afin que le quantificateur introduisant le voyageur porte sur toute la phrase et que les propriétés du chemin ne deviennent pas des propriétés du voyageur, fût-il fictif. Cette analyse sémantique (ou plutôt sa traduction en lambda-DRT) est d'ores et déjà implantée pour une partie du lexique de Grail.

Cet article traite de l'analyse de débats politiques selon une orientation multimodale. Nous étudions plus particulièrement les réponses aux interruptions lors d'un débat à l'Assemblée nationale. Nous proposons de procéder à l'analyse via des annotations systématiques de différentes modalités. L'analyse argumentative nous a amenée à proposer une typologie de ces réponses. Celle-ci a été mise à l'épreuve d'une classification automatique. La difficulté dans la construction d'un tel système réside dans la nature même des données : multimodales, parfois manquantes et incertaines.

La recherche d'entités nommées a été le sujet de nombreux travaux. Cependant, la construction des ressources nécessaires à de tels systèmes reste un problème majeur. Dans ce papier, nous proposons une méthode complémentaire aux outils capables de reconnaître des entités de types larges, dont l'objectif est de déterminer si une entité est d'un type donné, et ce de manière non-supervisée et quel que soit le type. Nous proposons pour cela une approche basée sur la comparaison de modèles de langage estimés à partir du Web. L'intérêt de notre approche est validé

par une évaluation sur 100 entités et 273 types différents.

En nous appuyant sur des données fournies par le concordancier bilingue TransSearch qui intègre un alignement statistique au niveau des mots, nous avons effectué une annotation semi-manuelle de la traduction anglaise de deux connecteurs du français. Les résultats de cette annotation montrent que les traductions de ces connecteurs ne correspondent pas aux « transpots » identifiés par TransSearch et encore moins à ce qui est proposé dans les dictionnaires bilingues.

Cet article présente l'utilisation d'un corpus comparable pour l'extraction de patrons de paraphrases. Nous présentons une méthode empirique basée sur l'appariement de n-grammes, permettant d'extraire des patrons de paraphrases dans des corpus comparables d'une même langue (le français), du même domaine (la médecine) mais de registres de langues différents (spécialisé ou grand public). Cette méthode confirme les résultats précédents basés sur des méthodes à base de patrons, et permet d'identifier de nouveaux patrons, apportant également un regard nouveau sur les différences entre les discours de langue générale et spécialisée.

Dans cet article, nous présentons une méthode de détection des correspondances bilingues de sous-catégorisation verbale à partir de données lexicales monolingues. Nous évoquons également la structure de ces lexiques et leur utilisation en traduction automatique (TA) à base linguistique anglais-japonais. Les lexiques sont utilisés par un programme de TA fonctionnant selon une architecture classique dite "à transfert", et leur structure permet une classification précise des sous-catégorisations verbales. Nos travaux ont permis une amélioration des données de sous-catégorisation des lexiques pour les verbes japonais et leurs équivalents anglais, en utilisant des données linguistiques compilées à partir d'un corpus de textes extrait du web. De plus, le fonctionnement du programme de TA a pu être amélioré en utilisant ces données.

Nous proposons dans cet article une méthode non-supervisée d'extraction des relations entre entités nommées. La méthode proposée se caractérise par l'utilisation de résultats d'analyses syntaxiques, notamment les chemins syntaxiques reliant deux entités nommées dans des arbres de dépendance. Nous avons également exploité la dualité de la représentation des relations sémantiques et le résultat de notre expérience comparative a montré que cette approche améliorait les rappels.

Ce travail prend place dans le cadre plus général du développement d'une plate-forme d'analyse syntaxique du français parlé. Nous décrivons la conception d'un modèle automatique pour résoudre le lien anaphorique présent dans les dislocations à gauche dans un corpus de français parlé radiophonique. La détection de ces structures devrait permettre à terme d'améliorer notre analyseur syntaxique en enrichissant les informations prises en compte dans nos modèles automatiques. La résolution du lien anaphorique est réalisée en deux étapes : un premier niveau à base de règles filtre les configurations candidates, et un second niveau s'appuie sur un modèle appris selon le critère du maximum d'entropie. Une évaluation expérimentale réalisée par validation croisée sur un corpus annoté manuellement donne une F-mesure de l'ordre de 40%.

Les familles de mots produites par deux analyseurs morphologiques, DériF (basé sur des règles) et Morphonette (basé sur l'analogie), appliqués à un même corpus lexical, sont comparées. Cette comparaison conduit à l'examen de trois sous-ensembles :
- un sous-ensemble commun aux deux systèmes dont la taille montre que, malgré leurs différences, les approches expérimentées par chaque système sont valides et décrivent en partie la même réalité morphologique.
- un sous-ensemble propre à DériF et un autre à Morphonette. Ces ensembles (a) nous renseignent sur les caractéristiques propres à chaque système, et notamment sur ce que l'autre ne peut pas produire, (b) ils mettent en évidence les erreurs d'un système, en ce qu'elles n'apparaissent pas

dans l'autre, (c) ils font apparaître certaines limites de la description, notamment celles qui sont liées aux objets et aux notions théoriques comme les familles morphologiques, les bases, l'existence de RCL « transversales » entre les lexèmes qui n'ont pas de relation d'ascendance ou de descendance.

Dans cet article nous présentons une série d'adaptations de l'algorithme du "cadre d'apprenstissage guidé" pour résoudre différentes tâches d'étiquetage. La spécificité du système proposé réside dans sa capacité à apprendre l'ordre de l'inférence avec les paramètres du classifieur local au lieu de la forcer dans un ordre pré-défini (de gauche à droite). L'algorithme d'entraînement est basé sur l'algorithme du "perceptron". Nous appliquons le système à différents types de tâches d'étiquetage pour atteindre des résultats au niveau de l'état de l'art en un court temps d'exécution.

Le travail présente une méthode de navigation dans les textes, fondée sur la répétition lexicale. La méthode choisie est celle développée par le linguiste Hoey. Son application manuelle à des textes de grandeur conséquente est problématique. Nous proposons dans cet article un processus automatique qui permet d'analyser selon cette méthode des textes de grande taille ; des expériences ont été menées appliquant le processus à différents types de textes (narratif, expositif) et montrant l'intérêt de l'approche.

L'exploitation de corpus analysés syntaxiquement (ou corpus arborés) pour le public non spécialiste n'est pas un problème trivial. Si la communauté du TAL souhaite mettre à la disposition des chercheurs non-informaticiens des corpus comportant des annotations linguistiques complexes, elle doit impérativement développer des interfaces simples à manipuler mais permettant des recherches fines. Dans cette communication, nous présentons les modes de recherche « grand public » développé(e)s dans le cadre du projet Scientext, qui met à disposition un corpus d'écrits scientifiques interrogeable par partie textuelle, par partie du discours et par fonction syntaxique. Les

modes simples sont décrits : un mode libre et guidé, où l'utilisateur sélectionne lui-même les éléments de la requête, et un mode sémantique, qui comporte des grammaires locales préétablies à l'aide des fonctions syntaxiques.

Cet article décrit deux expériences sur la construction de ressources terminologiques multi-lingues (preterminologies) préliminaires, mais grandes, grâce à des communautés Internet, et s'appuie sur ces expériences pour cibler des données terminologiques plus raffinées venant de communautés Internet et d'applications Web 2.0. La première expérience est une passerelle de contribution pour le site Web de la Route de la Soie numérique (DSR). Les visiteurs contribuent en effet à un référentiel lexical multi-lingue dédié, pendant qu'ils visitent et lisent les livres archivés, parce qu'ils sont intéressés par le domaine et ont tendance à être polygottes. Nous avons recueilli 1400 contributions lexicales en 4 mois. La seconde expérience est basée sur le JeuxDeMots arabe, où les joueurs en ligne contribuent à un réseau lexical arabe. L'expérience a entraîné une croissance régulière du nombre de joueurs et de contributions, ces dernières contenant des termes absents et des mots de dialectes oraux.

G-LexAr est un analyseur morphologique de l'arabe qui a récemment reçu des améliorations substantielles. Cet article propose une évaluation de cet analyseur en tant qu'outil de pré-traitement pour la traduction automatique statistique, ce dont il n'a encore jamais fait l'objet. Nous étudions l'impact des différentes formes proposées par son analyse (voyellation, lemmatisation et segmentation) sur un système de traduction arabe-anglais, ainsi que l'impact de la combinaison de ces formes. Nos expériences montrent que l'utilisation séparée de chacune de ces formes n'a que peu d'influence sur la qualité des traductions obtenues, tandis que leur combinaison y contribue de façon très bénéfique.

La projection de patrons lexico-syntaxiques sur corpus est une des manières privilégiées pour

identifier des relations sémantiques précises entre éléments lexicaux. Dans cet article, nous proposons d'étendre la notion de patron en prenant en compte la sémantique que véhiculent les éléments de structure d'un document (définitions, titres, énumérations) dans l'identification de relations. Nous avons testé cette hypothèse dans le cadre de la construction d'ontologies à partir de textes fortement structurés du domaine de la cartographie.

Dans cet article, nous discutons la méthodologie utilisée par Its-2, un système de traduction à base de règles, pour la traduction des pronoms clitiques. En particulier, nous nous focalisons sur les séquences clitiques, pour la traduction automatique entre le français et l'anglais. Une évaluation basée sur un corpus de phrases construites montre le potentiel de notre approche pour des traductions de bonne qualité.

Ce travail décrit la distribution des pronoms selon le style de texte (littéraire ou journalistique) et selon la langue (français, anglais, allemand et italien). Sur la base d'un étiquetage morpho-syntaxique effectué automatiquement puis vérifié manuellement, nous pouvons constater que la proportion des différents types de pronoms varie selon le type de texte et selon la langue. Nous discutons les catégories les plus ambiguës de manière détaillée. Comme nous avons utilisé l'analyseur syntaxique Fips pour l'étiquetage des pronoms, nous l'avons également évalué et obtenu une précision moyenne de plus de 95%.

Dans cette étude, notre système de traduction automatique, Its-2, a fait l'objet d'une évaluation manuelle de la traduction des pronoms pour cinq paires de langues et sur deux corpus : un corpus littéraire et un corpus de communiqués de presse. Les résultats montrent que les pourcentages d'erreurs peuvent atteindre 60% selon la paire de langues et le corpus. Nous discutons ainsi deux pistes de recherche pour l'amélioration des performances de Its-2 : la résolution des ambiguïtés d'analyse et la résolution des anaphores pronominales.

On étudie environ 500 occurrences du verbe « quitter » en les classant selon les inférences qu'elles suggèrent au lecteur. On obtient ainsi 43 « schémas inférentiels ». Ils ne s'excluent pas l'un l'autre : si plusieurs d'entre eux s'appliquent, les inférences produites se cumulent ; cependant, comme l'auteur sait que le lecteur dispose de tels schémas, s'il veut l'orienter vers une seule interprétation, il fournit des indices permettant d'éliminer les autres. On conjecture que ces schémas présentent des régularités observables sur des familles de mots, que ces régularités proviennent du fonctionnement d'opérations génériques, et qu'il est donc sans gravité de ne pas être exhaustif, dans la mesure où ces opérations permettent d'engendrer les schémas manquants en cas de besoin.

Dans cet article, nous présentons un système de détection d'opinions construit à partir des sorties d'un analyseur syntaxique robuste produisant des analyses profondes. L'objectif de ce système est l'extraction d'opinions associées à des produits (les concepts principaux) ainsi qu'aux concepts qui leurs sont associés (en anglais «features-based opinion extraction»). Suite à une étude d'un corpus cible, notre analyseur syntaxique est enrichi par l'ajout de polarité aux éléments pertinents du lexique et par le développement de règles génériques et spécialisées permettant l'extraction de relations sémantiques d'opinions, qui visent à alimenter un modèle de représentation des opinions. Une première évaluation montre des résultats très encourageants, mais de nombreuses perspectives restent à explorer.

Cet article décrit la première version et les résultats de l'évaluation d'un système de détection des épisodes d'infections associées aux soins. Cette détection est basée sur l'analyse automatique de comptes-rendus d'hospitalisation provenant de différents hôpitaux et différents services. Ces comptes-rendus sont sous forme de texte libre. Le système de détection a été développé à partir d'un analyseur linguistique que nous avons adapté au domaine médical et extrait à partir des

documents des indices pouvant conduire à une suspicion d'infection. Un traitement de la négation et un traitement temporel des textes sont effectués permettant de restreindre et de raffiner l'extraction d'indices. Nous décrivons dans cet article le système que nous avons développé et donnons les résultats d'une évaluation préliminaire.

La démonstration présentée produit une analyse syntaxique du français. Elle est écrite en SYGMART, fournie avec les actes, exécutable à l'adresse : http ://www.lirmm.fr/ chauche/ExempleAnl.html et téléchargeable à l'adresse : http ://www.sygtext.fr.

Nous présentons un système pour la classification en continu de dépêches financières selon une polarité positive ou négative. La démonstration permettra ainsi d'observer quelles sont les dépêches les plus à même de faire varier la valeur d'actions cotées en bourse, au moment même de la démonstration. Le système traitera de dépêches écrites en anglais et en français.

L'article décrit la structure et les applications possibles de la théorie des K-représentations (représentation des connaissances) dans la bioinformatique afin de développer un Réseau Sémantique d'une génération nouvelle. La théorie des K-répresentations est une théorie originale du développement des analyseurs sémantico?syntactiques avec l'utilisation large des moyens formels pour décrire les données d'entrée, intermédiaires et de sortie. Cette théorie est décrit dans la monographie de V. Fomichov (Springer, 2010). La première partie de la théorie est un modèle formel d'un système qui est composé de dix opérations sur les structures conceptuelles. Ce modèle définit une classe nouvelle des langages formels ? la classe des SK-langages. Les possibilités larges de construire des répresentations sémantiques des discours compliqués en rapport à la biologie sont manifestes. Une approche formelle nouvelle de l'élaboration des analysateurs mult-ilinguistiques sémantico-syntactiques est décrite. Cet approche a été implémentée sous la forme d'un programme en langage PYTHON.

Nous décrivons la création d'un environnement web pour aider des apprenants (adolescents ou adultes) à acquérir les automatismes nécessaires pour produire à un débit "normal" les structures fondamentales d'une langue. Notre point de départ est une base de données de phrases, glanées sur le web ou issues de livres scolaires ou de livres de phrases. Ces phrases ont été généralisées (remplacement de mots par des variables) et indexées en termes de buts pour former une arborescence de patrons. Ces deux astuces permettent de motiver l'usage des patrons et de créér des phrases structurellement identiques à celles rencontrées, tout en étant sémantiquement différentes. Si les notions de 'patrons' ou de 'phrases à trou implicitement typées' ne sont pas nouvelles, le fait de les avoir portées sur ordinateur pour apprendre des langues l'est. Le système étant conçu pour être ouvert, il permet aux utilisateurs, concepteurs ou apprenants, des changements sur de nombreux points importants : le nom des variables, leurs valeurs, le laps de temps entre une question et sa réponse, etc. La version initiale a été développée pour l'anglais et le japonais. Pour tester la généricité de notre approche nous y avons ajouté relativement facilement le français et le chinois.

Les encyclopédies numériques contiennent aujourd'hui de vastes inventaires de formes d'écritures pour des noms de personnes, de lieux, de produits ou d'organisation. Nous présentons un système hybride de détection d'entités nommées qui combine un classifieur à base de Champs Conditionnel Aléatoires avec un ensemble de motifs de détection extraits automatiquement d'un contenu encyclopédique. Nous proposons d'extraire depuis des éditions en plusieurs langues de l'encyclopédie Wikipédia de grandes quantités de formes d'écriture que nous utilisons en tant que motifs de détection des entités nommées. Nous décrivons une méthode qui nous assure de ne conserver dans cette ressources que des formes non ambiguës susceptibles de venir renforcer un système de détection d'entités nommées automatique. Nous procédons à un ensemble d'expériences qui nous permettent de comparer un système d'étiquetage à base de CRF avec un

système utilisant exclusivement des motifs de détection. Puis nous fusionnons les résultats des deux systèmes et montrons qu'un gain de performances est obtenu grâce à cette proposition.

Le titrage automatique de documents textuels est une tâche essentielle pour plusieurs applications (titrage de mails, génération automatique de sommaires, synthèse de documents, etc.). Cette étude présente une méthode de construction de titres courts appliquée à un corpus d'articles journalistiques via des méthodes de Fouille du Web. Il s'agit d'une première étape cruciale dans le but de proposer une méthode de construction de titres plus complexes. Dans cet article, nous présentons une méthode proposant des titres tenant compte de leur cohérence par rapport au texte, par rapport au Web, ainsi que de leur contexte dynamique. L'évaluation de notre approche indique que nos titres construits automatiquement sont informatifs et/ou accrocheurs.

Les systèmes d'extraction d'information traditionnels se focalisent sur un domaine spécifique et un nombre limité de relations. Les travaux récents dans ce domaine ont cependant vu émerger la problématique des systèmes d'extraction d'information à large échelle. À l'instar des systèmes de question-réponse en domaine ouvert, ces systèmes se caractérisent à la fois par le traitement d'un grand nombre de relations et par une absence de restriction quant aux domaines abordés. Dans cet article, nous présentons un système d'extraction d'information à large échelle fondé sur un apprentissage faiblement supervisé de patrons d'extraction de relations. Cet apprentissage repose sur la donnée de couples d'entités en relation dont la projection dans un corpus de référence permet de constituer la base d'exemples de relations support de l'induction des patrons d'extraction. Nous présentons également les résultats de l'application de cette approche dans le cadre d'évaluation défini par la tâche KBP de l'évaluation TAC 2010.

Le résumé automatique cross-lingue consiste à générer un résumé rédigé dans une langue différente de celle utilisée dans les documents sources. Dans cet article, nous proposons une

approche de résumé automatique multi-document, basée sur une représentation par graphe, qui prend en compte des scores de qualité de traduction lors du processus de sélection des phrases. Nous évaluons notre méthode sur un sous-ensemble manuellement traduit des données utilisées lors de la campagne d'évaluation internationale DUC 2004. Les résultats expérimentaux indiquent que notre approche permet d'améliorer la lisibilité des résumés générés, sans pour autant dégrader leur informativité.

Pourtant essentiel pour appréhender rapidement et globalement l'état de santé des patients, l'accès aux informations médicales liées aux prescriptions médicamenteuses et aux concepts médicaux par les outils informatiques se révèle particulièrement difficile. Ces informations sont en effet généralement rédigées en texte libre dans les comptes rendus hospitaliers et nécessitent le développement de techniques dédiées. Cet article présente les stratégies mises en oeuvre pour extraire les prescriptions médicales et les concepts médicaux dans des comptes rendus hospitaliers rédigés en anglais. Nos systèmes, fondés sur des approches à base de règles et d'apprentissage automatique, obtiennent une F1-mesure globale de 0,773 dans l'extraction des prescriptions médicales et dans le repérage et le typage des concepts médicaux.

Dans cet article, nous proposons plusieurs approches pour la portabilité du module de compréhension de la parole (SLU) d'un système de dialogue d'une langue vers une autre. On montre que l'utilisation des traductions automatiques statistiques (SMT) aide à réduire le temps et le cout de la portabilité d'un tel système d'une langue source vers une langue cible. Pour la tache d'étiquetage sémantique on propose d'utiliser soit les champs aléatoires conditionnels (CRF), soit l'approche à base de séquences (PH-SMT). Les résultats expérimentaux montrent l'efficacité des méthodes proposées pour une portabilité rapide du SLU vers une nouvelle langue. On propose aussi deux méthodes pour accroître la robustesse du SLU aux erreurs de traduction. Enfin on montre que la combinaison de ces approches réduit les erreurs du système. Ces travaux sont

motivés par la disponibilité du corpus MEDIA français et de la traduction manuelle vers l'italien d'une sous partie de ce corpus.

La fouille de données orales est un domaine de recherche visant à caractériser un flux audio contenant de la parole d'un ou plusieurs locuteurs, à l'aide de descripteurs liés à la forme et au contenu du signal. Outre la transcription automatique en mots des paroles prononcées, des informations sur le type de flux audio traité ainsi que sur le rôle et l'identité des locuteurs sont également cruciales pour permettre des requêtes complexes telles que : « chercher des débats sur le thème X », « trouver toutes les interviews de Y », etc. Dans ce cadre, et en traitant des conversations enregistrées lors d'émissions de radio ou de télévision, nous étudions la manière dont les locuteurs expriment des questions dans les conversations, en partant de l'intuition initiale que la forme des questions posées est une signature du rôle du locuteur dans la conversation (présentateur, invité, auditeur, etc.). En proposant une classification du type des questions et en utilisant ces informations en complément des descripteurs généralement utilisés dans la littérature pour classer les locuteurs par rôle, nous espérons améliorer l'étape de classification, et valider par la même occasion notre intuition initiale.

A rebours de bon nombre d'applications actuelles offrant des services de recherche d'information selon des critères temporels - applications qui reposent, à y regarder de près, sur une approche consistant à filtrer les résultats en fonction de leur inclusion dans une fenêtre de temps, nous souhaitons illustrer dans cet article l'intérêt d'un service s'appuyant sur un calcul de similarité entre des expressions adverbiales calendaires. Nous décrivons une heuristique pour mesurer la pertinence d'un fragment de texte en prenant en compte la sémantique des expressions calendaires qui y sont présentes. A travers la mise en oeuvre d'un système de recherche d'information, nous montrons comment il est possible de tirer profit de l'indexation d'expressions calendaires présentes dans les textes en définissant des scores de pertinence par rapport à une requête. L'objectif est de

faciliter la recherche d'information en offrant la possibilité de croiser des critères de recherche thématique avec des critères temporels.

La variabilité des corpus constitue un problème majeur pour les systèmes de reconnaissance d'entités nommées. L'une des pistes possibles pour y remédier est l'utilisation d'approches linguistiques pour les adapter à de nouveaux contextes : la construction de patrons sémantiques peut permettre de désambiguïser les entités nommées en structurant leur environnement syntaxico-sémantique. Cet article présente une première réalisation sur un corpus de presse d'un système de correction. Après une étape de segmentation sur des critères discursifs de surface, le système extrait et pondère les patrons liés à une classe d'entité nommée fournie par un analyseur. Malgré des modèles encore relativement élémentaires, les résultats obtenus sont encourageants et montrent la nécessité d'un traitement plus approfondi de la classe Organisation.

Effectuer une tâche de désambiguïsation lexicale peut permettre d'améliorer de nombreuses applications du traitement automatique des langues comme l'extraction d'informations multi-lingues, ou la traduction automatique. Schématiquement, il s'agit de choisir quel est le sens le plus approprié pour chaque mot d'un texte. Une des approches classiques consiste à estimer la proximité sémantique qui existe entre deux sens de mots puis de l'étendre à l'ensemble du texte. La méthode la plus directe donne un score à toutes les paires de sens de mots puis choisit la chaîne de sens qui a le meilleur score. La complexité de cet algorithme est exponentielle et le contexte qu'il est calculatoirement possible d'utiliser s'en trouve réduit. Il ne s'agit donc pas d'une solution viable. Dans cet article, nous nous intéressons à une autre méthode, l'adaptation d'un algorithme à colonies de fourmis. Nous présentons ses caractéristiques et montrons qu'il permet de propager à un niveau global les résultats des algorithmes locaux et de tenir compte d'un contexte plus long et plus approprié en un temps raisonnable.

Cet article est une prise de position concernant les plate-formes de type Amazon Mechanical Turk, dont l'utilisation est en plein essor depuis quelques années dans le traitement automatique des langues. Ces plateformes de travail en ligne permettent, selon le discours qui prévaut dans les articles du domaine, de faire développer toutes sortes de ressources linguistiques de qualité, pour un prix imbattable et en un temps très réduit, par des gens pour qui il s'agit d'un passe-temps. Nous allons ici démontrer que la situation est loin d'être aussi idéale, que ce soit sur le plan de la qualité, du prix, du statut des travailleurs ou de l'éthique. Nous rappellerons ensuite les solutions alternatives déjà existantes ou proposées. Notre but est ici double : informer les chercheurs, afin qu'ils fassent leur choix en toute connaissance de cause, et proposer des solutions pratiques et organisationnelles pour améliorer le développement de nouvelles ressources linguistiques en limitant les risques de dérives éthiques et légales, sans que cela se fasse au prix de leur coût ou de leur qualité.

Nous étudions dans cet article le problème de la comparabilité des documents composant un corpus comparable afin d'améliorer la qualité des lexiques bilingues extraits et les performances des systèmes de recherche d'information interlingue. Nous proposons une nouvelle approche qui permet de garantir un certain degré de comparabilité et d'homogénéité du corpus tout en préservant une grande part du vocabulaire du corpus d'origine. Nos expériences montrent que les lexiques bilingues que nous obtenons sont d'une meilleure qualité que ceux obtenus avec les approches précédentes, et qu'ils peuvent être utilisés pour améliorer significativement les systèmes de recherche d'information interlingue.

De nombreuses méthodes automatiques de classification de textes selon les sentiments qui y sont exprimés s'appuient sur un lexique dans lequel à chaque entrée est associée une valence. Le plus souvent, ce lexique est construit à partir d'un petit nombre de mots, choisis arbitrairement, qui servent de germes pour déterminer automatiquement la valence d'autres mots. La question de

l'optimalité de ces mots germes a bien peu retenu l'attention. Sur la base de la comparaison de cinq méthodes automatiques de construction de lexiques de valence, dont une qui, à notre connaissance, n'a jamais été adaptée au français et une autre développée spécifiquement pour la présente étude, nous montrons l'importance du choix de ces mots germes et l'intérêt de les identifier au moyen d'une procédure d'apprentissage supervisée.

Dans (Muller & Langlais, 2010), nous avons comparé une approche distributionnelle et une variante de l'approche miroir proposée par Dyvik (2002) sur une tâche d'extraction de synonymes à partir d'un corpus en français. Nous présentons ici une analyse plus fine des relations extraites automatiquement en nous intéressant cette fois-ci à la langue anglaise pour laquelle de plus amples ressources sont disponibles. Différentes façons d'évaluer notre approche corroborent le fait que l'approche miroir se comporte globalement mieux que l'approche distributionnelle décrite dans (Lin, 1998), une approche de référence dans le domaine.

L'alignement et la mesure d'accord sur des textes multi-annotés sont des enjeux majeurs pour la constitution de corpus de référence. Nous défendons dans cet article l'idée que ces deux tâches sont par essence interdépendantes, la mesure d'accord nécessitant de s'appuyer sur des annotations alignées, tandis que les choix d'alignements ne peuvent se faire qu'à l'aune de la mesure qu'ils induisent. Nous proposons des principes formels relevant cette gageure, qui s'appuient notamment sur la notion de désordre du système constitué par l'ensemble des jeux d'annotations d'un texte. Nous posons que le meilleur alignement est celui qui minimise ce désordre, et que la valeur de désordre obtenue rend compte simultanément du taux d'accord. Cette approche, qualifiée d'holiste car prenant en compte l'intégralité du système pour opérer, est algorithmiquement lourde, mais nous sommes parvenus à produire une implémentation d'une version légèrement dégradée de cette dernière, et l'avons intégrée à la plate-forme d'annotation Glozz.

Cet article a un double objectif : d'une part, il s'agit de présenter à la communauté un corpus récemment rendu public, le French Time Bank (FTiB), qui consiste en une collection de textes journalistiques annotés pour les temps et les événements selon la norme ISO-TimeML ; d'autre part, nous souhaitons livrer les résultats et réflexions méthodologiques que nous avons pu tirer de la réalisation de ce corpus de référence, avec l'idée que notre expérience pourra s'avérer profitable au-delà de la communauté intéressée par le traitement de la temporalité.

Les applications de recherche d'informations chez Orange sont confrontées à des flux importants de données textuelles, recouvrant des domaines larges et évoluant très rapidement. Un des problèmes à résoudre est de pouvoir analyser très rapidement ces flux, à un niveau élevé de qualité. Le recours à un modèle d'analyse sémantique, comme solution, n'est viable qu'en s'appuyant sur l'apprentissage automatique pour construire des grandes bases de connaissances dédiées à chaque application. L'extraction terminologique décrite dans cet article est un composant amont de ce dispositif d'apprentissage. Des nouvelles méthodes d'acquisition, basée sur un modèle hybride (analyse par grammaires de chunking et analyse statistique à deux niveaux), ont été développées pour répondre aux contraintes de performance et de qualité.

Nous présentons dans cet article une nouvelle manière d'aborder le problème de l'acquisition automatique de paires de mots en relation de traduction à partir de corpus comparables. Nous décrivons tout d'abord les approches standard et par similarité interlangue traditionnellement dédiées à cette tâche. Nous réinterprétons ensuite la méthode par similarité interlangue et motivons un nouveau modèle pour reformuler cette approche inspirée par les métamoteurs de recherche d'information. Les résultats empiriques que nous obtenons montrent que les performances de notre modèle sont toujours supérieures à celles obtenues avec l'approche par similarité interlangue, mais aussi comme étant compétitives par rapport à l'approche standard.

Depuis septembre 2007, un réseau lexical de grande taille pour le Français est en cours de construction à l'aide de méthodes fondées sur des formes de consensus populaire obtenu via des jeux (projet JeuxDeMots). L'intervention d'experts humains est marginale en ce qu'elle représente moins de 0,5% des relations du réseau et se limite à des corrections, à des ajustements ainsi qu'à la validation des sens de termes. Pour évaluer la qualité de cette ressource construite par des participants de jeu (utilisateurs non experts) nous adoptons une démarche similaire à celle de sa construction, à savoir, la ressource doit être validée sur un vocabulaire de classe ouverte, par des non-experts, de façon stable (persistante dans le temps). Pour ce faire, nous proposons de vérifier si notre ressource est capable de servir de support à la résolution du problème nommé 'Mot sur le Bout de la Langue' (MBL). A l'instar de JeuxdeMots, l'outil développé peut être vu comme un jeu en ligne. Tout comme ce dernier, il permet d'acquérir de nouvelles relations, constituant ainsi un enrichissement de notre réseau lexical.

L'identification de la cible d'une d'opinion fait l'objet d'une attention récente en fouille d'opinion. Les méthodes existantes ont été testées sur des corpus monothématiques en anglais. Elles permettent principalement de traiter les cas où la cible se situe dans la même phrase que l'opinion. Dans cet article, nous abordons cette problématique pour le français dans un corpus multithématique et nous présentons une nouvelle méthode pour identifier la cible d'une opinion apparaissant hors du contexte phrastique. L'évaluation de la méthode montre une amélioration des résultats par rapport à l'existant.

Dans cet article, nous synthétisons les résultats de plusieurs séries d'expériences réalisées à l'aide de CRF (Conditional Random Fields ou "champs markoviens conditionnels") linéaires pour apprendre à annoter des textes français à partir d'exemples, en exploitant diverses ressources linguistiques externes. Ces expériences ont porté sur l'étiquetage morpho-syntaxique intégrant

l'identification des unités polylexicales. Nous montrons que le modèle des CRF est capable d'intégrer des ressources lexicales riches en unités multi-mots de différentes manières et permet d'atteindre ainsi le meilleur taux de correction d'étiquetage actuel pour le français.

Pour la plupart des langues utilisant l'alphabet latin, le découpage d'un texte selon les espaces et les symboles de ponctuation est une bonne approximation d'un découpage en unités lexicales. Bien que cette approximation cache de nombreuses difficultés, elles sont sans comparaison avec celles que l'on rencontre lorsque l'on veut traiter des langues qui, comme le chinois mandarin, n'utilisent pas l'espace. Un grand nombre de systèmes de segmentation ont été proposés parmi lesquels certains adoptent une approche non-supervisée motivée linguistiquement. Cependant les méthodes d'évaluation communément utilisées ne rendent pas compte de toutes les propriétés de tels systèmes. Dans cet article, nous montrons qu'un modèle simple qui repose sur une reformulation en termes d'entropie d'une hypothèse indépendante de la langue énoncée par Harris (1955), permet de segmenter un corpus et d'en extraire un lexique. Testé sur le corpus de l'Academia Sinica, notre système permet l'induction d'une segmentation et d'un lexique qui ont de bonnes propriétés intrinsèques et dont les caractéristiques sont similaires à celles du lexique sous-jacent au corpus segmenté manuellement. De plus, on constate une certaine corrélation entre les résultats du modèle de segmentation et les structures syntaxiques fournies par une sous-partie arborée corpus.

Les connaissances morphologiques sont fréquemment utilisées en Question-Réponse afin de faciliter l'appariement entre mots de la question et mots du passage contenant la réponse. Il n'existe toutefois pas d'étude qualitative et quantitative sur les phénomènes morphologiques les plus pertinents pour ce cadre applicatif. Dans cet article, nous présentons une analyse détaillée des phénomènes de morphologie constructionnelle permettant de faire le lien entre question et réponse. Pour ce faire, nous avons constitué et annoté un corpus de paires de questions-réponses, qui nous a permis de construire une ressource de référence, utile pour l'évaluation de la couverture de

ressources et d'outils d'analyse morphologique. Nous détaillons en particulier les phénomènes de dérivation et de composition et montrons qu'il reste un nombre important de relations morphologiques dérivationnelles pour lesquelles il n'existe pas encore de ressource exploitable pour le français.

Nous montrons dans une série d'expériences sur quatre langues, sur des échantillons du corpus Europarl, que, dans leur grande majorité, les tri-grammes inconnus d'un jeu de test peuvent être reconstruits par analogie avec des tri-grammes hapax du corpus d'entraînement. De ce résultat, nous dérivons une méthode de lissage simple pour les modèles de langue par tri-grammes et obtenons de meilleurs résultats que les lissages de Witten-Bell, Good-Turing et Kneser-Ney dans des expériences menées en onze langues sur la partie commune d'Europarl, sauf pour le finnois et, dans une moindre mesure, le français.

Nous présentons une architecture pour l'analyse syntaxique en deux étapes. Dans un premier temps un analyseur syntagmatique construit, pour chaque phrase, une liste d'analyses qui sont converties en arbres de dépendances. Ces arbres sont ensuite réévalués par un réordonnanceur discriminant. Cette méthode permet de prendre en compte des informations auxquelles l'analyseur n'a pas accès, en particulier des annotations fonctionnelles. Nous validons notre approche par une évaluation sur le corpus arboré de Paris 7. La seconde étape permet d'améliorer significativement la qualité des analyses retournées, quelle que soit la métrique utilisée.

Dans cet article, nous nous intéressons à l'identification de relations entre entités en domaine de spécialité, et étudions l'apport d'informations syntaxiques. Nous nous plaçons dans le domaine médical, et analysons des relations entre concepts dans des comptes-rendus médicaux, tâche évaluée dans la campagne i2b2 en 2010. Les relations étant exprimées par des formulations très variées en langue, nous avons procédé à l'analyse des phrases en extrayant des traits qui

concourent à la reconnaissance de la présence d'une relation et nous avons considéré l'identification des relations comme une tâche de classification multi-classes, chaque catégorie de relation étant considérée comme une classe. Notre système de référence est celui qui a participé à la campagne i2b2, dont la F-mesure est d'environ 0,70. Nous avons évalué l'apport de la syntaxe pour cette tâche, tout d'abord en ajoutant des attributs syntaxiques à notre classifieur, puis en utilisant un apprentissage fondé sur la structure syntaxique des phrases (apprentissage à base de tree kernels) ; cette dernière méthode améliore les résultats de la classification de 3%.

Nous montrons comment enrichir une annotation en dépendances syntaxiques au format du French Treebank de Paris 7 en utilisant la réécriture de graphes, en vue du calcul de sa représentation sémantique. Le système de réécriture est composé de règles grammaticales et lexicales structurées en modules. Les règles lexicales utilisent une information de contrôle extraite du lexique des verbes français Dicovalence.

Les approches classiques à base de n-grammes en analyse supervisée de sentiments ne peuvent pas correctement identifier les expressions complexes de sentiments à cause de la perte d'information induite par l'approche « sac de mots » utilisée pour représenter les textes. Dans notre approche, nous avons recours à des sous-graphes extraits des graphes de dépendances syntaxiques comme traits pour la classification de sentiments. Nous représentons un texte par un vecteur composé de ces sous-graphes syntaxiques et nous employons un classifieurs SVM état-de-l'art pour identifier la polarité d'un texte. Nos évaluations expérimentales sur des critiques de jeux vidéo montrent que notre approche à base de sous-graphes est meilleure que les approches standard à modèles « sac de mots » et n-grammes. Dans cet article nous avons travaillé sur le français, mais notre approche peut facilement être adaptée à d'autres langues.

Nous montrons que les différentes interprétations d'une combinaison de plusieurs GN peuvent être

modélisées par deux opérations de combinaison sur les référents de ces GN, appelées combinaison cumulative et combinaison distributive. Nous étudions aussi bien les GN définis et indéfinis que les GN quantifiés ou pluriels et nous montrons comment la combinaison d'un GN avec d'autres éléments peut induire des interprétations collective ou individualisante. Selon la façon dont un GN se combine avec d'autres GN, le calcul de son référent peut être fonction de ces derniers ; ceci définit une relation d'ancrage de chaque GN, qui induit un ordre partiel sur les GN. Considérer cette relation plutôt que la relation converse de portée simplifie le calcul de l'interprétation des GN et des énoncés. Des représentations sémantiques graphiques et algébriques sans considération de la portée sont proposées pour les dites alternances de portée.

De nombreux phénomènes linguistiques visent à exprimer le doute ou l'incertitude de l'énonciateur, ainsi que la subjectivité potentielle du point de vue. La prise en compte de ces informations sur le niveau de certitude est primordiale pour de nombreuses applications du traitement automatique des langues, en particulier l'extraction d'information dans le domaine médical. Dans cet article, nous présentons deux systèmes qui analysent automatiquement les niveaux de certitude associés à des problèmes médicaux mentionnés dans des compte-rendus cliniques en anglais. Le premier système procède par apprentissage supervisé et obtient une f-mesure de 0,93. Le second système utilise des règles décrivant des déclencheurs linguistiques spécifiques et obtient une f-mesure de 0,90.

Les annotations discursives proposées dans le cadre de théories discursives comme RST (Rhetorical Structure Theory) ou SDRT (Segmented Dicourse Representation Theory) ont comme point fort de construire une structure discursive globale liant toutes les informations données dans un texte. Les annotations discursives proposées dans le PDTB (Penn Discourse Tree Bank) ont comme point fort d'identifier la "source" de chaque information du texte?répondant ainsi à la question qui a dit ou pense quoi ? Nous proposons une approche unifiée pour les annotations discursives alliant les points forts de ces deux courants de recherche. Cette approche unifiée repose

crucialement sur des information de factivité, telles que celles qui sont annotées dans le corpus (anglais) FactBank.

Dans cet article, nous analysons les modifications locales disponibles dans l'historique des révisions de la version française de Wikipédia. Nous définissons tout d'abord une typologie des modifications fondée sur une étude détaillée d'un large corpus de modifications. Puis, nous détaillons l'annotation manuelle d'une partie de ce corpus afin d'évaluer le degré de complexité de la tâche d'identification automatique de paraphrases dans ce genre de corpus. Enfin, nous évaluons un outil d'identification de paraphrases à base de règles sur un sous-ensemble de notre corpus.

Dans ce document, nous présentons les principales caractéristiques de <TextCoop>, un environnement basé sur les grammaires logiques dédié à l'analyse de structures discursives. Nous étudions en particulier le langage DisLog qui fixe la structure des règles et des spécifications qui les accompagnent. Nous présentons la structure du moteur de <TextCoop> en indiquant au fur et à mesure du texte l'état du travail, les performances et les orientations en particulier en matière d'environnement, d'aide à l'écriture de règles et de développement applicatif.

Dans cet article, nous présentons une analyse à base de contraintes de la relation forme-sens des gestes déictiques et de leur signal de parole synchrone. En nous basant sur une étude empirique de corpus multimodaux, nous définissons quels énoncés multimodaux sont bien formés, et lesquels ne pourraient jamais produire le sens voulu dans la situation communicative. Plus précisément, nous formulons une grammaire multimodale dont les règles de construction utilisent la prosodie, la syntaxe et la sémantique de la parole, la forme et le sens du signal déictique, ainsi que la performance temporelle de la parole et la deixis afin de contraindre la production d'un arbre de syntaxe combinant parole et gesture déictique ainsi que la représentation unifiée du sens pour l'action multimodale correspondant à cet arbre. La contribution de notre projet est double : nous

ajoutons aux ressources existantes pour le TAL un corpus annoté de parole et de gestes, et nous créons un cadre théorique pour la grammaire au sein duquel la composition sémantique d'un énoncé découle de la synchronie entre geste et parole.

L'alignement sous-phrastique consiste à extraire des traductions d'unités textuelles de grain inférieur à la phrase à partir de textes multi-lingues parallèles alignés au niveau de la phrase. Un tel alignement est nécessaire, par exemple, pour entraîner des systèmes de traduction statistique. L'approche standard pour réaliser cette tâche implique l'estimation successive de plusieurs modèles probabilistes de complexité croissante et l'utilisation d'heuristiques qui permettent d'aligner des mots isolés, puis, par extension, des groupes de mots. Dans cet article, nous considérons une approche alternative, initialement proposée dans (Lardilleux & Lepage, 2008), qui repose sur un principe beaucoup plus simple, à savoir la comparaison des profils d'occurrences dans des sous-corpus obtenus par échantillonnage. Après avoir analysé les forces et faiblesses de cette approche, nous montrons comment améliorer la détection d'unités de traduction longues, et évaluons ces améliorations sur des tâches de traduction automatique.

Dans les systèmes de traduction statistique à base de segments, le modèle de traduction est estimé à partir d'alignements mot-à-mot grâce à des heuristiques d'extraction et de valuation. Bien que ces alignements mot-à-mot soient construits par des modèles probabilistes, les processus d'extraction et de valuation utilisent ces modèles en faisant l'hypothèse que ces alignements sont déterministes. Dans cet article, nous proposons de lever cette hypothèse en considérant l'ensemble de la matrice d'alignement, d'une paire de phrases, chaque association étant valuée par sa probabilité. En comparaison avec les travaux antérieurs, nous montrons qu'en utilisant un modèle exponentiel pour estimer de manière discriminante ces probabilités, il est possible d'obtenir des améliorations significatives des performances de traduction. Ces améliorations sont mesurées à l'aide de la métrique BLEU sur la tâche de traduction de l'arabe vers l'anglais de l'évaluation NIST MT'09, en

considérant deux types de conditions selon la taille du corpus de données parallèles utilisées.

Dans cet article, nous décrivons une nouvelle méthode d'alignement automatique de paraphrases d'énoncés. Nous utilisons des méthodes développées précédemment afin de produire différentes approches hybrides (hybridations). Ces différentes méthodes permettent d'acquérir des équivalences textuelles à partir d'un corpus monolingue parallèle. L'hybridation combine des informations obtenues par diverses techniques : alignements statistiques, approche symbolique, fusion d'arbres syntaxiques et alignement basé sur des distances d'édition. Nous avons évalué l'ensemble de ces résultats et nous constatons une amélioration sur l'acquisition de paraphrases sous-phrastiques.

Dans cet article, nous présentons un système de segmentation non supervisée que nous évaluons sur des données en mandarin. Notre travail s'inspire de l'hypothèse de Harris (1955) et suit Kempe (1999) et Tanaka-Ishii (2005) en se basant sur la reformulation de l'hypothèse en termes de variation de l'entropie de branchement. Celle-ci se révèle être un bon indicateur des frontières des unités linguistiques. Nous améliorons le système de (Jin et Tanaka-Ishii, 2006) en ajoutant une étape de normalisation qui nous permet de reformuler la façon dont sont prises les décisions de segmentation en ayant recours à la programmation dynamique. Ceci nous permet de supprimer la plupart des seuils de leur modèle tout en obtenant de meilleurs résultats, qui se placent au niveau de l'état de l'art (Wang et al., 2011) avec un système plus simple que ces derniers. Nous présentons une évaluation des résultats sur plusieurs corpus diffusés pour le Chinese Word Segmentation bake-off II (Emerson, 2005) et détaillons la borne supérieure que l'on peut espérer atteindre avec une méthode non-supervisée. Pour cela nous utilisons ZPAR en apprentissage croisé (Zhang et Clark, 2010) comme suggéré dans (Huang et Zhao, 2007; Zhao et Kit, 2008)

Nous proposons une étude dynamique du lexique, en décrivant la manière dont il s'organise

progressivement du début à la fin d'un texte. Pour ce faire, nous nous focalisons sur la co-occurrence généralisée, en formant un graphe qui représente tous les lemmes du texte et synthétise leurs relations mutuelles de co-occurrence. L'étude d'un corpus de 40 textes montre que ces relations évoluent d'une manière auto-organisée : la forme - et l'identité - du graphe de co-occurrence restent stables après une phase d'organisation terminée avant la 1ère moitié du texte. Ensuite, il n'évolue plus : les nouveaux mots et les nouvelles relations de co-occurrence s'inscrivent peu à peu dans le réseau, sans modifier la forme d'ensemble de la structure. La relation de co-occurrence généralisée dans un texte apparaît donc comme la construction rapide d'un système, qui est ensuite assez souple pour canaliser un flux d'information sans changer d'identité.

L'article évalue un éventail de mesures de similarité qui ont pour but de prédire les scores de similarité sémantique et les relations sémantiques qui s'établissent entre deux termes, et étudie les moyens de combiner ces mesures. Nous présentons une analyse comparative à grande échelle de 34 mesures basées sur des réseaux sémantiques, le Web, des corpus, ainsi que des définitions. L'article met en évidence les forces et les faiblesses de chaque approche en contexte de l'extraction de relations. Enfin, deux techniques de combinaison de mesures sont décrites et testées. Les résultats montrent que les mesures combinées sont plus performantes que toutes les mesures simples et aboutissent à une corrélation de 0,887 et une Precision(20) de 0,979.

Ce travail porte sur l'application d'une technique de traduction statistique au problème de la normalisation de textos. La méthode est basée sur l'algorithme de recherche vorace décrit dans (Langlais et al., 2007). Une première normalisation est générée, puis nous appliquons itérativement une fonction qui génère des nouvelles hypothèses à partir de la normalisation courante, et maximisons une fonction de score. Cette méthode fournit une réduction du taux d'erreurs moyen par phrase de 33 % sur le corpus de test, et une augmentation du score BLEU de plus de 30 %. Nous mettons l'accent sur les fonctions qui génèrent la normalisation initiale et sur les opérations

permettant de générer des nouvelles hypothèses.

Dans ce papier, nous introduisons le problème que pose la correction orthographique sur des corpus de qualité très dégradée tels que les messages publiés sur les forums, les sites d'avis ou les réseaux sociaux. Nous proposons une première architecture de correction qui a pour objectif d'éviter au maximum la sur-correction. Nous présentons, par ailleurs l'implémentation et les résultats d'un des modules de ce système qui a pour but de détecter si un mot inconnu, dans une phrase de langue connue, est un mot qui appartient à cette langue ou non.

Cet article présente et évalue une plate-forme ouverte et flexible pour l'acquisition automatique d'expressions polylexicales (EPL) à partir des corpus monolingues. Nous commençons par une motivation pratique suivie d'une discussion théorique sur le comportement et les défis posés par les EPL dans les applications de TAL. Ensuite, nous décrivons les modules de notre plate-forme, leur enchaînement et les choix d'implémentation. L'évaluation de la plate-forme a été effectuée à travers une applications : la lexicographie assistée par ordinateur. Cette dernière peut bénéficier de l'acquisition d'EPL puisque les expressions acquises automatiquement à partir des corpus peuvent à la fois accélérer la création et améliorer la qualité et la couverture des ressources lexicales. Les résultats prometteurs encouragent une recherche plus approfondie sur la manière optimale d'intégrer le traitement des EPL dans de nombreuses applications de TAL, notamment dans les systèmes traduction automatique.

Nous présentons ici un système de prédiction de néologismes formels avec pour exemple la génération automatique de néologismes nominaux suffixés par -IER dénotant des artefacts (saladier, brassière, thonier). L'objectif de cet article est double. Il s'agira (i) de mettre en évidence les contraintes de la suffixation par -IER afin de les implémenter dans un système de génération morphologique puis (ii) de montrer qu'il est possible de prédire les néologismes formels. Ce

système de prédiction permettrait ainsi de compléter automatiquement les lexiques pour le Traitement Automatique des Langues (TAL).

L'article présente une extension de l'analyseur traditionnel en dépendances par transitions adapté aux dépendances discontinues et les premiers résultats de son entraînement sur un corpus de structures de dépendances de phrases en français. Les résultats des premières expérimentations vont servir de base pour le choix des traits des configurations de calcul bien adaptés aux dépendances discontinues pour améliorer l'apprentissage des dépendances tête.

Dans cette étude, nous nous intéressons au problème de l'analyse d'un corpus annoté de l'oral. Le système d'annotation considéré est celui introduit par l'équipe des syntacticiens du projet Rhapsodie. La principale problématique qui sous-tend un tel projet est que la base écrite sur laquelle on travaille est en réalité une transcription de l'oral, balisée par les annotateurs de manière à délimiter un ensemble de structures arborescentes. Un tel système introduit plusieurs structures, en particulier macro et micro-syntaxiques. Du fait de leur étroite imbrication, il s'est avéré difficile de les analyser de façon indépendante et donc de travailler sur l'aspect macro-syntaxique indépendamment de l'aspect micro-syntaxique. Cependant, peu d'études jusqu'à présent considèrent ces problèmes conjointement et de manière automatisée. Dans ce travail, nous présentons nos efforts en vue de produire un outil de parsing capable de rendre compte à la fois de l'information micro et macro-syntaxique du texte annoté. Pour ce faire, nous proposons une représentation partant de la notion de multi-arbre et nous montrons comment une telle structure peut être générée à partir de l'annotation et utilisée à des fins d'analyse.

La recherche présentée 1 s'inscrit dans le domaine de la fouille d'opinion, domaine qui consiste principalement à déterminer la polarité d'un texte ou d'une phrase. Dans cette optique, le contexte autour d'un mot polarisé joue un rôle essentiel, car il peut modifier la polarité initiale de ce terme.

Nous avons choisi d'approfondir cette question et de détecter précisément ces modifieurs de polarité. Une étude exploratoire, décrite dans des travaux antérieurs, nous a permis d'extraire automatiquement des adverbes qui jouent un rôle sur la polarité des adjectifs auxquels ils sont associés et de préciser leur impact. Nous avons ensuite amélioré le système d'extraction afin de construire automatiquement un lexique de structures lexico-syntaxiques modifiantes associées au type d'impact qu'elles ont sur un terme polarisé. Nous présentons ici le fonctionnement du système actuel ainsi que l'évaluation du lexique obtenu.

Ce travail s'inscrit dans le cadre de l'analyse morphologique et syntaxique automatique de la langue arabe. Nous nous intéressons au traitement de la diacritisation et à son apport pour l'analyse morphologique. En effet, la plupart des analyseurs morphologiques et des étiqueteurs morpho-syntaxiques existants ignorent les diacritiques présents dans le texte à analyser et commettent des erreurs qui pourraient être évitées. Dans cet article, nous proposons une méthode qui prend en considération les diacritiques lors de l'analyse, et nous montrons que cette prise en compte permet de diminuer considérablement le taux d'erreur de l'analyse morphologique selon le taux de diacritiques du texte traité.

Dans le cadre d'apprentissages humains assistés par des environnements informatiques, les techniques de TAL ne sont que rarement employées ou restreintes à des tâches ou des domaines spécifiques comme l'ALAO (Apprentissage de la Langue Assisté par Ordinateur) où elles sont omniprésentes mais ne concernent que certaines dimensions du TAL. Nous cherchons à explorer les possibilités ou les performances des techniques voire des méthodes de TAL pour des systèmes moins spécifiques dès lors qu'une dimension de réseau et de collectivité est présente. Plus particulièrement, notre objectif est d'obtenir des indicateurs sur la construction collective de connaissances, et ses modalités. Ce papier présente la problématique de notre thèse, son contexte, nos motivations ainsi que nos premières réflexions.

Dans cet article, nous présentons les principales méthodes non supervisées à base de connaissances pour la désambiguïsation lexicale. Elles sont composées d'une part de mesures de similarité sémantique locales qui donnent une valeur de proximité entre deux sens de mots et, d'autre part, d'algorithmes globaux qui utilisent les mesures de similarité sémantique locales pour trouver les sens appropriés des mots selon le contexte à l'échelle de la phrase ou du texte.

La présente recherche cherche à réduire la taille de messages textuels sur la base de techniques de compression observées, pour la plupart, dans un corpus de sms. Ce papier explique la méthodologie suivie pour établir des règles de contraction. Il présente ensuite les 33 règles retenues, et illustre les quatre niveaux de compression proposés par deux exemples concrets, produits automatiquement par un premier prototype. Le but de cette recherche n'est donc pas de produire de "l'écrit-sms", mais d'élaborer un procédé de compression capable de produire des textes courts et compréhensibles à partir de n'importe quelle source textuelle en français. Le terme "d'essentialisation" est proposé pour désigner cette approche de réduction textuelle.

Cet article présente une approche permettant de reconnaitre automatiquement dans un texte des séquences verbales figées (casser sa pipe, briser la glace, prendre en compte) à partir d'une ressource. Cette ressource décrit chaque séquence en termes de possibilités et de restrictions transformationnelles. En effet, les séquences figées ne le sont pas complètement et nécessitent une description exhaustive afin de ne pas extraire seulement les formes canoniques. Dans un premier temps nous aborderons les approches traditionnelles permettant d'extraire des séquences phraséologiques. Par la suite, nous expliquerons comment est constituée notre ressource et comment celle-ci est utilisée pour un traitement automatique.

Les travaux sur l'émotion dans les forums sont nombreux en Linguistique et Psychologie. L'objectif

de cette contribution est de proposer une analyse de l'émotion dans les forums de santé selon l'angle des Sciences de l'Information et de la Communication mais également selon une approche interdisciplinaire. Il s'agira ici, d'étudier l'émotion comme un critère de pertinence lorsque des personnes malades effectuent des recherches dans les forums. Ce papier introduit la méthodologie utilisée en traitement automatique de la langue afin de répondre à cette interrogation. Ainsi, le travail présenté abordera l'exploitation d'un corpus de messages de forums, la catégorisation semi-supervisée et l'utilisation du logiciel NooJ pour traiter de manière automatique les données.

Nous présentons une approche de peuplement d'ontologie dont le but est de modéliser le comportement de composants logiciels afin de faciliter le passage de descriptions d'exigences en langue naturelle à des spécifications formelles. L'ontologie que nous cherchons à peupler a été conçue à partir des connaissances du domaine de la domotique et est initialisée à partir d'une description de la configuration physique d'un environnement intelligent. Notre méthode est guidée par l'extraction d'instances de relations permettant par là-même d'extraire les instances de concepts liés par ces relations. Nous construisons des règles d'extraction à partir d'éléments issus de l'analyse syntaxique de descriptions de besoins utilisateurs et de ressources terminologiques associées aux concepts et relations de l'ontologie. Notre approche de peuplement se distingue par sa finalité qui n'est pas d'extraire toutes les instances décrivant un domaine mais d'extraire des instances pouvant participer sans conflit à un des multiples fonctionnements décrit par des utilisateurs.

Une des principales faiblesses des jeux sérieux à l'heure actuelle est qu'ils incorporent très souvent des questionnaires à choix multiple (QCM). Or, aucune étude n'a démontré que les QCM sont capables d'évaluer précisément le niveau de compréhension des apprenants. Au contraire, certaines études ont montré expérimentalement que permettre à l'apprenant d'entrer une phrase libre dans le programme au lieu de simplement cocher une réponse dans un QCM rend possible

une évaluation beaucoup plus fine des compétences de l'apprenant. Nous proposons donc de concevoir un agent conversationnel capable de comprendre des énoncés en langage naturel dans un cadre sémantique restreint, cadre correspondant au domaine de compétence testé chez l'apprenant. Cette fonctionnalité est destinée à permettre un dialogue naturel avec l'apprenant, en particulier dans le cadre des jeux sérieux. Une telle interaction en langage naturel a pour but de masquer les QCM sous-jacents. Cet article présente notre approche.

Comme le E-commerce est devenu de plus en plus populaire, le nombre de commentaires des internautes est en croissance constante. Les opinions sur le Web affectent nos choix et nos décisions. Il s'avère alors indispensable de traiter une quantité importante de critiques des clients afin de présenter à l'utilisateur l'information dont il a besoin dans la forme la plus appropriée. Dans cet article, nous présentons ResTS, un nouveau système de résumé automatique de textes d'opinions basé sur les caractéristiques des produits. Notre approche vise à transformer les critiques des utilisateurs en des scores qui mesurent le degré de satisfaction des clients pour un produit donné et pour chacune de ses caractéristiques. Ces scores sont compris entre 0 et 1 et peuvent être utilisés pour la prise de décision. Nous avons étudié les opinions véhiculées par les noms, les adjectifs, les verbes et les adverbes, contrairement aux recherches précédentes qui utilisent essentiellement les adjectifs. Les résultats expérimentaux préliminaires montrent que notre méthode est comparable aux méthodes classiques de résumé automatique basées sur les caractéristiques des produits.

L'évaluation des systèmes de question-réponse lors des campagnes repose généralement sur la validité d'une réponse individuelle supportée par un passage (question factuelle) ou d'un groupe de réponses toutes contenues dans un même passage (questions listes). Ce cadre évaluatif empêche donc de fournir un ensemble de plusieurs réponses individuelles et ne permet également pas de fournir des réponses provenant de documents différents. Ce recoupement inter-documents peut

être necessaire pour construire une réponse composée de plusieurs éléments afin d'être le plus complet possible. De plus une grande majorité de questions formulées au singulier et semblant n'attendre qu'une seule réponse se trouve être des questions possédant plusieurs réponses correctes. Nous présentons ici une typologie des questions à réponses multiples ainsi qu'un aperçu sur les problèmes posés à un système de question-réponse par ce type de question.

La portabilité entre les langues des systèmes de reconnaissance d'entités nommées est coûteuse en termes de temps et de connaissances linguistiques requises. L'adaptation des systèmes symboliques souffrent du coût de développement de nouveaux lexiques et de la mise à jour des règles contextuelles. D'un autre côté, l'adaptation des systèmes statistiques se heurtent au problème du coût de préparation d'un nouveau corpus d'apprentissage. Cet article étudie l'intérêt et le coût associé pour porter un système existant de reconnaissance d'entités nommées pour du texte bien formé vers une autre langue. Nous présentons une méthode peu coûteuse pour porter un système symbolique dédié au français vers l'anglais. Pour ce faire, nous avons d'une part traduit automatiquement l'ensemble des lexiques de mots déclencheurs au moyen d'un dictionnaire bilingue. D'autre part, nous avons manuellement modifié quelques règles de manière à respecter la syntaxe de la langue anglaise. Les résultats expérimentaux sont comparés à ceux obtenus avec un système de référence développé pour l'anglais.

En langue des signes, l'espace est utilisé pour localiser et faire référence à certaines entités dont l'emplacement est important pour la compréhension du sens. Dans cet article, nous proposons une représentation informatique de l'espace de signation et les fonctions de création et d'accès associées, afin d'analyser les gestes manuels et non manuels qui contribuent à la localisation et au référencement des signes et de matérialiser leur effet. Nous proposons une approche bi-directionnelle qui se base sur l'analyse de données de capture de mouvement de discours en langue des signes dans le but de caractériser les événements de localisation et de référencement.

L'intérêt pour la fouille d'opinion s'est développé en même temps que se sont répandus les blogs, forums et autres plate-formes où les internautes peuvent librement exprimer leur opinion. La très grande quantité de données disponibles oblige à avoir recours à des traitements automatiques de fouille d'opinion. Cependant, la manière dont les gens expriment leur avis change selon ce dont ils parlent. Les distributions des mots utilisés sont différentes d'un domaine à l'autre. Aussi, il est très difficile d'obtenir un classifieur d'opinion fonctionnant sur tous les domaines. De plus, on ne peut appliquer sans adaptation sur un domaine cible un classifieur entraîné sur un domaine source différent. L'objet de cet article est de recenser les moyens de résoudre ce problème difficile.

L'accès au contenu des textes de spécialité est une tâche difficile à réaliser. Cela nécessite la définition de méthodes automatiques ou semi-automatiques pour identifier des relations sémantiques entre les termes que contiennent ces textes. Nous distinguons les approches de TAL permettant d'acquérir ces relations suivant deux types d'information : la structure interne des termes ou le contexte de ces termes en corpus. Afin d'améliorer la qualité des relations acquises et faciliter leur réutilisation en corpus, nous nous intéressons à la prise en compte du contexte dans une méthode d'acquisition de relations de synonymie basée sur l'utilisation de la structure interne des termes. Nous présentons les résultats d'une expérience préliminaire tenant compte de l'usage des termes dans un corpus biomédical en anglais. Nous donnons quelques pistes de travail pour définir des contraintes sémantiques sur les relations de synonymie acquises.

Cet article explique la chaîne de traitement suivie pour extraire une grammaire PCFG à partir du corpus de Paris VII. Dans un premier temps cela nécessite de transformer les arbres syntaxiques du corpus en arbres de dérivation d'une grammaire AB, ce que nous effectuons en utilisant un transducteur d'arbres généralisé ; il faut ensuite extraire de ces arbres une PCFG. Le transducteur d'arbres généralisé est une variation des transducteurs d'arbres classiques et c'est l'extraction de la

grammaire à partir des arbres de dérivation qui donnera l'aspect probabiliste à la grammaire. La PCFG extraite est utilisée via l'algorithme CYK pour l'analyse de phrases.

Cet article présente une plateforme open-source pour l'édition collaborative de corpus de dépendances. Cette plateforme, nommée ACOLAD (Annotation de COrpus Linguistique pour l'Analyse de Dépendances), propose des services manuels de segmentation et d'annotation multi-niveaux (segmentation en mots et en syntagmes minimaux (chunks), annotation morphosyntaxique des mots, annotation syntaxique des chunks et annotation syntaxique des dépendances entre mots ou entre chunks). Dans cet article, nous présentons la plateforme ACOLAD, puis nous détaillons la représentation pivot utilisée pour gérer les annotations concurrentes, enfin décrivons le mécanisme d'importation de ressources linguistiques externes.

Cet article présente une approche linguistique pour l'extraction d'expressions de préférence à partir de dialogues de négociation. Nous proposons un nouveau schéma d'annotation pour encoder les préférences et les dépendances exprimées linguistiquement dans deux genres de corpus différents. Ensuite, nous proposons une méthode d'apprentissage qui extrait les expressions de préférence en utilisant une combinaison de traits locaux et discursifs. Finalement, nous évaluons la fiabilité de notre approche sur chaque genre de corpus.

La présence de conflits dans les communautés épistémiques en ligne peut s'avérer bloquante pour l'activité de conception. Nous présentons une étude sur la détection automatique de conflit dans les discussions entre contributeurs Wikipedia qui s'appuie sur des traits de surface tels que la subjectivité ou la connotation des énoncés et évaluons deux règles de décision : l'une découle d'un modèle dialectique en exploitant localement la structure linéaire de la discussion, la subjectivité et la connotation ; l'autre, plus globale, ne s'appuie que sur la taille des fils et les marques de subjectivité au détriment des marques de connotation. Nous montrons que ces deux règles produisent des

résultats similaires mais que la simplicité de la règle globale en fait une approche préférée dans la détection des conflits.

Les travaux liés à la définition et à la reconnaissance des entités nommées sont généralement envisagés en domaine ouvert, à travers la conception de catégories génériques (noms de personnes, de lieux, etc.) et leur application à des données textuelles issues de la presse (orale comme écrite). Par ailleurs, la fouille des données issues de centres d'appel est stratégique pour une entreprise comme EDF, compte tenu du rôle crucial joué par l'opinion pour les applications marketing, ce qui passe par la définition d'entités d'intérêt propres au domaine. Nous comparons les deux types de modèles d'entités - génériques et spécifiques à un domaine précis - afin d'observer leurs points de recouvrement, via l'annotation manuelle d'un corpus de conversations en centres d'appel. Nous souhaitons ainsi étudier l'apport d'une détection en entités nommées génériques pour l'extraction d'information métier en domaine restreint.

Depuis une dizaine d'années, le TAL s'intéresse à la subjectivité, notamment dans la perspective d'applications telles que la fouille d'opinion et l'analyse des sentiments. Or, la linguistique de corpus outillée par des méthodes textométriques a souvent abordé la question de la subjectivité dans les textes. Notre objectif est de montrer d'une part, ce que pourrait apporter à l'analyse des sentiments l'analyse textométrique et d'autre part, comment mutualiser les avantages d'une association entre celle-ci et une méthode de classification automatique basée sur l'apprentissage supervisé. En nous appuyant sur un corpus de témoignages issus de forums de discussion, nous montrerons que la prise en compte de critères sélectionnés suivant une analyse textométrique permet d'obtenir des résultats de classification satisfaisants par rapport à une vision purement lexicale.

Cet article présente une méthodologie visant, à partir d'une observation de corpus vidéo de langue des signes, à repérer puis formaliser les régularités de structure dans les constructions

linguistiques. Cette méthodologie est applicable à tous les niveaux du langage, du sub-lexical à l'énoncé complet. En s'appuyant sur deux exemples, il présente une application de cette méthodologie ainsi que le modèle AZee qui, intégrant la souplesse nécessaire en termes de synchronisation des articulateurs, permet une formalisation des règles repérées.

Cet article présente une campagne d'annotation de commentaires de matchs de football en français. L'annotation a été réalisée à partir d'un corpus très hétérogène, contenant à la fois des comptes-rendus minute par minute et des transcriptions des commentaires vidéo. Nous montrons ici comment les accords intra- et inter-annotateurs peuvent être utilisés efficacement, en en proposant une définition adaptée à notre type de tâche et en mettant en exergue l'importance de certaines bonnes pratiques concernant leur utilisation. Nous montrons également comment certains indices collectés à l'aide d'outils statistiques simples peuvent être utilisés pour indiquer des pistes de corrections des annotations. Ces différentes propositions nous permettent par ailleurs d'évaluer l'impact des modalités sources de nos textes (oral ou écrit) sur le coût et la qualité des annotations.

Le travail présenté dans cet article se situe dans le contexte de la fouille d'opinion et se focalise sur la détermination de la polarité d'un texte en adoptant une approche par apprentissage. Dans ce cadre, son objet est d'étudier différentes stratégies d'adaptation à un nouveau domaine dans le cas de figure fréquent où des données d'entraînement n'existent que pour un ou plusieurs domaines différents du domaine cible. Cette étude montre en particulier que l'utilisation d'une forme d'auto-apprentissage par laquelle un classifieur annote un corpus du domaine cible et modifie son corpus d'entraînement en y incorporant les textes classés avec la plus grande confiance se révèle comme la stratégie la plus performante et la plus stable pour les différents domaines testés. Cette stratégie s'avère même supérieure dans un nombre significatif de cas à la méthode proposée par (Blitzer et al., 2007) sur les mêmes jeux de test tout en étant plus simple.

Dans le cadre du projet franco-allemand Emolex, dédié à l'étude contrastive de la combinatoire du lexique des émotions en 5 langues, nous avons développé des outils et des méthodes permettant l'extraction, la visualisation et la comparaison de profls combinatoires pour des expressions simples et complexes. Nous présentons ici l'architecture d'ensemble de la plate-forme, conçue pour efectuer des extractions sur des corpus de grandes dimensions (de l'ordre de la centaine de millions de mots) avec des temps de réponse réduits (le corpus étant interrogeable en ligne1). Nous décrivons comment nous avons introduit la notion de pivots complexes, afn de permettre aux utilisateurs de rafner progressivement leurs requêtes pour caractériser des constructions lexico-syntaxiques élaborées. Enfn, nous donnons les premiers résultats d'un module d'extraction automatique d'expressions polylexicales récurrentes.

L'approche standard d'identification d'équivalents terminologiques à partir de corpus comparables repose sur la comparaison de mots contextuels en langues source et cible et sur l'utilisation d'un lexique bilingue. Nous analysons manuellement, selon des critères linguistiques (parties du discours, spécificité et relations sémantiques), les propriétés des mots contextuels et des erreurs commises par l'approche standard appliquée à la terminologie médicale pour suggérer des améliorations basées sur la sélection de mots contextuels.

Nous décrivons BiTermEx, un prototype d'expérimentation de l'extraction de terminologie bilingue de mots composés, à partir de documents comparables, via la méthode compositionnelle. Nous expliquons la variation morphologique et la combinaison des constituants lexicaux des termes composés. Cette permet une précision TOP1 de 92% et 97,5% en français anglais, et de 94% en français japonais pour l'alignement de termes composés (textes scientifiques et de vulgarisation scientifique).

Dans cet article, nous présentons un système d'extraction automatique d'événements fondé sur

deux approches actuelles en extraction d'information : la première s'appuie sur des règles linguistiques construites manuellement et la seconde se fonde sur un apprentissage automatique de patrons linguistiques. Les expérimentations réalisées montrent que combiner ces deux méthodes d'extraction permet d'améliorer significativement la qualité des événements extraits (amélioration de près de 10 points de F-mesure).

Nous décrivons dans cet article comment nous avons procédé pour apprendre automatiquement un chunker à partir du French Tree Bank, en utilisant les CRF (Conditional Random Fields). Nous avons réalisé diverses expériences, pour reconnaître soit l'ensemble de tous les chunks possibles, soit les seuls groupes nominaux simples. Nous évaluons le chunker obtenu aussi bien de manière interne (sur le French Tree Bank lui-même) qu'externe (sur un corpus distinct transcrit de l'oral), afin de mesurer sa robustesse.

Cet article étudie le lien entre la similarité textuelle et une classification extrinsèque dans des collections de rapports d'incidents aéronautiques. Nous cherchons à compléter les stratégies d'analyse de ces collections en établissant automatiquement des liens de similarité entre les documents de façon à ce qu'ils ne reflètent pas l'organisation des schémas de codification utilisés pour leur classement. Afin de mettre en évidence les dimensions de variation transversales à la classification, nous calculons un score de dépendance entre les termes et les classes et excluons du calcul de similarité les termes les plus corrélés à une classe donnée. Nous montrons par une application sur 500 documents que cette méthode permet effectivement de dégager des thématiques qui seraient passées inaperçues au vu de la trop grande saillance des similarités de haut niveau.

Les performances des systèmes de traduction automatique statistique dépendent de la disponibilité de textes parallèles bilingues, appelés aussi bitextes. Cependant, les corpus parallèles sont des

ressources limitées et parfois indisponibles pour certains couples de langues ou domaines. Nous présentons une technique pour l'extraction de phrases parallèles à partir d'un corpus comparable multimodal (audio et texte). Ces enregistrements sont transcrits avec un système de reconnaissance automatique de la parole et traduits avec un système de traduction automatique. Ces traductions sont ensuite utilisées comme requêtes d'un système de recherche d'information pour sélectionner des phrases parallèles sans erreur et générer un bitexte. Plusieurs expériences ont été menées sur les données de la campagne IWSLT'11 (TED) qui montrent la faisabilité de notre approche.

La résolution d'anaphores est l'une des tâches les plus difficiles du Traitement Automatique du Langage Naturel (TALN). La capacité de classifier les pronoms avant de tenter une tâche de résolution d'anaphores serait importante, puisque pour traiter un pronom cataphorique le système doit chercher l'antécédent dans le segment qui suit le pronom. Alors que, pour le pronom anaphorique, le système doit chercher l'antécédent dans le segment qui précède le pronom. En outre, le nombre des pronoms a été jugée non-trivial dans la langue arabe. C'est dans ce cadre que se situe notre travail qui consiste à proposer une méthode pour la classification automatique des pronoms démonstratifs arabes, basée sur l'apprentissage. Nous avons évalué notre approche sur un corpus composé de 365585 mots contenant 14318 pronoms démonstratifs et nous avons obtenu des résultats encourageants : 99.3% comme F-Mesure.

Dans cet article, nous présentons un outil d'extraction et de normalisation d'un sous-ensemble d'expressions temporelles développé pour le français. Cet outil est mis au point et utilisé dans le cadre du projet ANR Chronolines1 et il est appliqué sur un corpus fourni par l'AFP. Notre but final dans le cadre du projet est de construire semi-automatiquement des chronologies événementielles à partir de la base de depêches de l'AFP. L'une des étapes du traitement est l'analyse de l'information temporelle véhiculée dans les textes. Nous avons donc développé un annotateur

d'expressions temporelles pour le français que nous décrivons dans cet article. Nous présenterons également les résultats de son évaluation.

Nous présentons les premiers pas vers la création d'un corpus annoté en discours pour le français : le French Discourse TreeBank enrichissant le FTB. La méthodologie adoptée s'inspire du Penn Discourse TreeBank (PDTB) mais elle s'en distingue sur au moins deux points à caractère théorique. D'abord, notre objectif est de fournir une couverture totale d'un texte du corpus, tandis que le PDTB ne fournit qu'une couverture partielle, qui ne peut donc pas être qualifiée d'analyse discursive comme celle faite en RST ou SDRT, deux théories majeures sur le discours. Ensuite, nous avons été amenés à définir une nouvelle hiérarchie des relations de discours qui s'inspire de RST, de SDRT et du PDTB.

L'utilisation de sources externes d'informations pour la recherche documentaire a été considérablement étudiée dans le passé. Des améliorations de performances ont été mises en lumière avec des corpus larges ou structurés. Néanmoins, dans ces études les ressources sont souvent utilisées séparément mais rarement combinées. Nous présentons une évaluation de la combinaison de quatre différentes ressources générales, standards et accessibles. Nous utilisons une mesure de distance informative pour extraire les caractéristiques contextuelles des différentes ressources et améliorer la représentation de la requête. Cette évaluation est menée sur une tâche de recherche d'information sur le Web en utilisant le corpus ClueWeb09 et les topics de la piste Web de TREC. Les meilleurs résultats sont obtenus en combinant les quatre ressources, et sont statistiquement significativement supérieurs aux autres approches.

La détection des Entités Nommées (EN) en langue arabe est un prétraitement potentiellement utile pour de nombreuses applications du traitement des langues, en particulier pour la traduction automatique. Cette tâche représente toutefois un sérieux défi, compte tenu des spécificités de

l'arabe. Dans cet article, nous présentons un compte-rendu de nos efforts pour développer un système de repérage des EN s'appuyant sur des méthodes statistiques, en détaillant les aspects liés à la sélection des caractéristiques les plus utiles pour la tâche ; puis diverses tentatives pour adapter ce système d'une manière entièrement non supervisée.

Les ressources lexicales sont cruciales pour de nombreuses applications de traitement automatique de la langue (par exemple, l'extraction d'opinions à partir de corpus). Cependant, leur construction pose des problèmes à différents niveaux (coût, couverture, etc.). Dans cet article, nous avons voulu vérifier si les informations morphologiques liées à la dérivation pouvaient être exploitées pour l'annotation automatique d'informations sémantiques. En partant d'une ressource regroupant les mots en familles morphologiques en français, nous avons construit un lexique de polarités pour 4 065 mots, à partir d'une liste initiale d'adjectifs annotés manuellement. Les résultats obtenus montrent que la propagation des polarités est correcte pour 7-8,89% des familles avec un seul adjectif. Le lexique ainsi obtenu améliore aussi les résultats du système d'extraction d'opinions.

Le travail présenté dans cet article est centré sur la constitution d'un corpus de textes journalistiques annotés au niveau discursif d'un point de vue thématique. Le modèle d'annotation est une segmentation classique, à laquelle nous ajoutons un repérage de zones de transition entre unités thématiques. Nous faisons l'hypothèse que dans un texte bien construit, le scripteur fournit des indications aidant le lecteur à passer d'un sujet à un autre, l'identification de ces indices étant susceptible d'améliorer les procédures de segmentation automatique. Les annotations produites ont fait l'objet d'analyses quantitatives mettant en évidence un ensemble de propriétés des transitions entre thèmes.

A partir de l'évaluation d'extracteurs de termes menée initialement pour détecter le meilleur outil d'acquisition du lexique d'une langue contrôlée, nous proposons dans cet article une stratégie

d'optimisation du processus d'extraction terminologique. Nos travaux, menés dans le cadre du projet ANR Sensunique, prouvent que la « multiextraction », c'est-à-dire la coopération de plusieurs extracteurs de termes, donne des résultats significativement meilleurs que l'extraction via un seul outil. Elle permet à la fois de réduire le silence et de filtrer automatiquement le bruit grâce à la variation d'un indice relatif au potentiel terminologique.

Dans cet article, nous nous sommes intéressés à la prise en compte des erreurs dans les contenus textuels des documents XML. Nous proposons une approche visant à diminuer l'impact de ces erreurs sur les systèmes de Recherche d'Information (RI). En effet, ces systèmes produisent des index associant chaque document aux termes qu'il contient. Les erreurs affectent donc la qualité des index ce qui conduit par exemple à considérer à tort des documents mal indexés comme non pertinents (resp. pertinents) vis-à-vis de certaines requêtes. Afin de faire face à ce problème, nous proposons d'inclure un mécanisme de correction d'erreurs lors de la phase d'indexation des documents. Nous avons implémenté cette approche au sein d'un prototype que nous avons évalué dans le cadre de la campagne d'évaluation INEX.

Cet article présente une approche de post-édition statistique pour adapter aux domaines de spécialité des systèmes de traduction automatique génériques. En utilisant les traductions produites par ces systèmes, alignées avec leur traduction de référence, un modèle de post-édition basé sur un alignement sous-phrastique est construit. Les expériences menées entre le français et l'anglais pour le domaine médical montrent qu'une telle adaptation a posteriori est possible. Deux systèmes de traduction statistiques sont étudiés : une implémentation locale état-de-l'art et un outil libre en ligne. Nous proposons aussi une méthode de sélection de phrases à post-éditer permettant d'emblée d'accroître la qualité des traductions et pour laquelle les scores oracles indiquent des gains encore possibles.

Le Corpus Arboré de Paris 7 (ou French TreeBank) est le corpus de référence pour le français aux niveaux morphosyntaxique et syntaxique. Toutefois, il ne contient pas d'annotations explicites en entités nommées. Ces dernières sont pourtant parmi les informations les plus utiles pour de nombreuses tâches en traitement automatique des langues et de nombreuses applications. De plus, aucun corpus du français annoté en entités nommées et de taille importante ne contient d'annotation référentielle, qui complète les informations de typage et d'empan sur chaque mention par l'indication de l'entité à laquelle elle réfère. Nous avons annoté manuellement avec ce type d'informations, après pré-annotation automatique, le Corpus Arboré de Paris 7. Nous décrivons les grandes lignes du guide d'annotation sous-jacent et nous donnons quelques informations quantitatives sur les annotations obtenues.

Cet article concerne des travaux effectués dans le cadre du 7ème atelier de traduction automatique statistique et du projet ANR COSMAT. Ces travaux se focalisent sur l'estimation de paramètres contenus dans une table de traduction. L'approche classique consiste à estimer ces paramètres à partir de fréquences relatives d'éléments de traduction. Dans notre approche, nous proposons d'utiliser le concept de masses de croyance afin d'estimer ces paramètres. La théorie des fonctions de croyances est une théorie très adaptée à la gestion des incertitudes dans de nombreux domaines. Les expériences basées sur notre approche s'appliquent sur la traduction de la paire de langue français-anglais dans les deux sens de traduction.

Cet article s'intéresse aux conversations téléphoniques d'un Centre d'Appels EDF, automatiquement découpées en « tours de parole » et automatiquement transcrites. Il fait apparaître une relation entre la longueur des tours de parole et leur contenu, en ce qui concerne le vocabulaire qui les compose et les sentiments qui y sont véhiculés. Après avoir montré qu'il y a un intérêt à étudier ces longs tours, l'article analyse leur contenu et liste quelques exemples autour des notions d'argumentation et de réclamation. Il montre ainsi que la longueur des tours de parole peut

être un critère utile de sélection de conversations.

L'ouverture du Centre National de Réception des Appels d'Urgence (CNRAU) accessible aux sourds et malentendants fait émerger des questions linguistiques qui portent sur le français écrit des sourds, et des questions informatiques dans le domaine du traitement automatique du langage naturel. Le français écrit des sourds, pratiqué par une population hétérogène, comporte des spécificités morpho-syntaxiques et morpho-lexicales qui peuvent rendre problématique la communication écrite entre les personnes sourdes appelantes et les agents du CNRAU. Un premier corpus de français écrit sourd élicité avec mise en situation d'urgence (FAX-ESSU) a été recueilli dans la perspective de proposer des solutions TAL et linguistiques aux agents du CNRAU dans le cadre de ces échanges écrits. Nous présentons une première étude lexicale, morpho-syntaxique et syntaxique de ce corpus reposant en partie sur une chaîne de traitement automatique, afin de valider les phénomènes linguistiques décrits dans la littérature et d'enrichir la connaissance du français écrit des sourds.

Nous présentons un outil de réécriture de graphes qui a été conçu spécifiquement pour des applications au TAL. Il permet de décrire des graphes dont les noeuds contiennent des structures de traits et dont les arcs décrivent des relations entre ces noeuds. Nous présentons ici la réécriture de graphes que l'on considère, l'implantation existante et quelques expérimentations.

Deux types de démonstrateurs sont présentés. Une première interface à visée didactique permet d'observer des traitements automatiques sur des documents vidéo. Plusieurs niveaux de représentation peuvent être montrés simultanément, ce qui facilite l'analyse d'approches multi-vues. La seconde interface est une interface opérationnelle de "consommation" de documents audio. Elle offre une expérience de navigation enrichie dans des documents audio grâce à une visualisation de méta-données extraites automatiquement.

Le logiciel Syntox, dont une interface utilisateur en ligne ce trouve à cette URL : http://www.syntox.net, est une mise en application d'un modèle basé sur les grammaires attribuées, dans le cadre de la synthèse de texte. L'outil est une plateforme d'expérimentation dont l'ergonomie est simple. Syntox est capable de traiter des lexiques et des grammaires volumineux sur des textes ambigus à partir de la description explicite de phénomènes linguistiques.

Nous proposons une démonstration de deux programmes : un segmenteur-étiqueteur POS pour le français et un programme de parenthésage en "chunks" de textes préalablement traités par le programme précédent. Tous deux ont été appris à partir du French Tree Bank.

SPPAS est le nouvel outil du LPL pour l'alignement texte/son. La segmentation s'opère en 4 étapes successives dans un processus entièrement automatique ou semi-automatique, à partir d'un fichier audio et d'une transcription. Le résultat comprend la segmentation en unités inter-pausales, en mots, en syllabes et en phonèmes. La version actuelle propose un ensemble de ressources qui permettent le traitement du français, de l'anglais, de l'italien et du chinois. L'ajout de nouvelles langues est facilitée par la simplicité de l'architecture de l'outil et le respect des formats de fichiers les plus usuels. L'outil bénéficie en outre d'une documentation en ligne et d'une interface graphique afin d'en faciliter l'accessibilité aux non-informaticiens. Enfin, SPPAS n'utilise et ne contient que des ressources et programmes sous licence libre GPL.

Proxem développe depuis 2007 une plate-forme de traitement du langage, Antelope, qui permet de construire rapidement des applications sémantiques verticales (par exemple, pour l'e-réputation, la veille économique ou l'analyse d'avis de consommateurs). Antelope a servi à créer une solution pour les Ressources Humaines, utilisée notamment par l'APEC, permettant (1) d'extraire de l'information à partir d'offres et de CVs et (2) de trouver les offres d'emploi correspondant le mieux à

un CV (ou réciproquement). Nous présentons ici l'adaptation d'Antelope à un domaine particulier, en l'occurrence les RH.

Cette démonstration présente NOMAO, un moteur de recherche géolocalisé qui permet à ses utilisateurs de trouver des lieux (bars, magasins...) qui correspondent à leurs goûts, à ceux de leurs amis et aux recommandations des internautes.

Le DictAm est un dictionnaire électronique des verbes amazighs-français. Il vise à rendre compte de l'ensemble des verbes dans le domaine berbère : conjugaison, diathèse et sens. Le DictAm comporte actuellement près de 3000 verbes dans une soixantaine de parlers berbères. C'est un travail qui est en cours de réalisation et qui a pour ambition de répertorier tous les verbes berbères ainsi que leurs équivalents en français.

L'objectif du système Vizart3D est de fournir à un locuteur, en temps réel, et de façon automatique, un retour visuel sur ses propres mouvements articulatoires. Les applications principales de ce système sont l'aide à l'apprentissage des langues étrangères et la rééducation orthophonique (correction phonétique). Le système Vizart3D est basé sur la tête parlante 3D développée au GIPSA-lab, qui laisse apparaître, en plus des lèvres, les articulateurs de la parole normalement cachés (comme la langue). Cette tête parlante est animée automatiquement à partir du signal audio de parole, à l'aide de techniques de conversion de voix et de régression acoustico-articulatoire par GMM.

ROCme! permet une gestion rationalisée, autonome et dématérialisée de l'enregistrement de corpus oraux. Il dispose notamment d'une interface pour le recueil de méta-données sur les locuteurs totalement paramétrable via des balises XML. Les locuteurs peuvent gérer les réponses au questionnaire, l'enregistrement audio, la lecture, la sauvegarde et le défilement des phrases (ou

autres types de corpus) en toute autonomie. ROCme! affiche du texte, avec ou sans mise en forme HTML, des images, du son et des vidéos.

L'extraction de relations par apprentissage nécessite un corpus annoté de très grande taille pour couvrir toutes les variations d'expressions des relations. Pour contrer ce problème, nous proposons une méthode de simplification de phrases qui permet de réduire la variabilité syntaxique des relations. Elle nécessite l'annotation d'un petit corpus qui sera par la suite augmenté automatiquement. La première étape est l'annotation des simplifications grâce à un classifieur à base de CRF, puis l'extraction des relations, et ensuite une complétion automatique du corpus d'entraînement des simplifications grâce aux résultats de l'extraction des relations. Les premiers résultats que nous avons obtenus pour la tâche d'extraction de relations d'i2b2 2010 sont très encourageants.

Cette recherche est issue de notre volonté de tester de nouvelles méthodes automatiques d'annotation ou d'extraction d'information à partir d'une langue L1 en exploitant des ressources et des outils disponibles pour une autre langue L2. Cette approche repose sur le passage par un corpus parallèle (L1-L2) aligné au niveau des phrases et des mots. Pour faire face au manque de corpus médicaux français annotés, nous nous intéressons au couple de langues (françaisanglais) dans le but d'annoter automatiquement des textes médicaux en français. En particulier, nous nous intéressons dans cet article à la reconnaissance des entités médicales. Nous évaluons dans un premier temps notre méthode de reconnaissance d'entités médicales sur le corpus anglais. Dans un second temps, nous évaluons la reconnaissance des entités médicales du corpus français par projection des annotations du corpus anglais. Nous abordons également le problème de l'hétérogénéité des données en exploitant un corpus extrait du Web et nous proposons une méthode statistique pour y pallier.

Dans les systèmes d'extraction d'information sur des événements, une tâche importante est le remplissage automatique de formulaires regroupant les informations sur un événement donné à partir d'un texte non structuré. Ce remplissage de formulaire peut s'avérer difficile lorsque l'information est dispersée dans tout le texte et mélangée à des éléments d'information liés à un autre événement similaire. Nous proposons dans cet article une approche en deux étapes pour ce problème : d'abord une segmentation du texte en événements pour sélectionner les phrases relatives au même événement ; puis une méthode de sélection dans les phrases sélectionnées des entités liées à l'événement. Une évaluation de cette approche sur un corpus annoté de dépêches dans le domaine des événements sismiques montre un F-score de 72% pour la tâche de remplissage de formulaires.

Cet article présente notre utilisation de la théorie des types dans laquelle nous nous situons pour l'analyse syntaxique, sémantique et pour la construction du lexique. Notre outil, Grail permet de traiter le discours automatiquement à partir du texte brut et nous le testons sur un corpus de récit de voyages pyrénéens, Itinerpy. Nous expliquons donc notre usage des grammaires catégorielles et plus particulièrement du calcul de Lambek et la correspondance entre ces catégories et le lambda-calcul simplement typé dans le cadre de la DRT. Une flexibilité du typage doit être autorisée dans certains cas et bloquée dans d'autres. Quelques phénomènes linguistiques participant à une forme de glissement de sens provocant des conflits de types sont présentés. Nous expliquons ensuite nos motivations d'ordre pragmatique à utiliser un système à sortes et types variables en sémantique lexicale puis notre traitement compositionnel du temps des évènements inspiré du Binary Tense de (Verkuyl, 2008).

Nous proposons deux stratégies discriminantes d'intégration des mots composés dans un processus réel d'analyse syntaxique : (i) pré-segmentation lexicale avant analyse, (ii) post-segmentation lexicale après analyse au moyen d'un réordonnanceur. Le segmenteur de

l'approche (i) se fonde sur un modèle CRF et permet d'obtenir un reconnaisseur de mots composés état-de-l'art. Le réordonnanceur de l'approche (ii) repose sur un modèle MaxEnt intégrant des traits dédiés aux mots composés. Nous montrons que les deux approches permettent de combler jusqu'à 18% de l'écart entre un analyseur baseline et un analyseur avec segmentation parfaite et jusqu'à 25% pour la reconnaissance des mots composés.

L'analyse syntaxique fine en dépendances nécessite la connaissance des cadres de souscatégorisation des unités lexicales. Le cas des verbes étant bien étudié, nous nous intéressons dans cet article au cas des noms communs dérivés de verbes. Notre intérêt principal est de calculer le cadre de sous-catégorisation des noms déverbaux à partir de celui du verbe d'origine pour le français. Or, pour ce faire il faut disposer d'une liste représentative de noms déverbaux français. Pour calculer cette liste nous utilisons un algorithme simplifié de repérage des noms déverbaux, l'appliquons à un corpus et comparons la liste obtenue avec la liste Verbaction des déverbaux exprimant l'action ou l'activité du verbe. Pour les noms déverbaux ainsi obtenus et attestés ensuite par une expertise linguistique, nous analysons la provenance des groupes prépositionnels subordonnés des déverbaux dans des contextes différents en tenant compte du verbe d'origine. L'analyse est effectuée sur le corpus Paris 7 et est limitée au cas le plus fréquent du génitif, c'est-à-dire des groupes prépositionnels introduits par de, des, etc.

Dans cette prise de position, nous nous intéressons au calcul de similarité (ou distances) entre textes, problématique présente dans de nombreuses tâches de TAL. Nous nous efforçons de montrer que ce qui n'est souvent qu'un composant dans des systèmes plus complexes est parfois négligé et des solutions sous-optimales sont employées. Ainsi, le calcul de similarité par TF-IDF/cosinus est souvent présenté comme « état-de-l'art », alors que des alternatives souvent plus performantes sont employées couramment dans le domaine de la Recherche d'Information (RI). Au travers de quelques expériences concernant plusieurs tâches, nous montrons combien ce

simple calcul de similarité peut influencer les performances d'un système. Nous considérons plus particulièrement deux alternatives. La première est le schéma de pondération Okapi-BM25, bien connu en RI et directement interchangeable avec le TF-IDF. L'autre, la vectorisation, est une technique de calcul de similarité que nous avons développée et qui offrent d'intéressantes propriétés.

Nous présentons dans cet article un travail portant sur la création d'un corpus de français parlé spontané annoté en morphosyntaxe. Nous détaillons la méthodologie suivie afin d'assurer le contrôle de la qualité de la ressource finale. Ce corpus est d'ores et déjà librement diffusé pour la recherche et peut servir aussi bien de corpus d'apprentissage pour des logiciels que de base pour des descriptions linguistiques. Nous présentons également les résultats obtenus par deux étiqueteurs morphosyntaxiques entrainés sur ce corpus.

Nous présentons un algorithme d'alignement sous-phrastique permettant d'aligner très facilement un couple de phrases à partir d'une matrice d'alignement pré-remplie. Cet algorithme s'inspire de travaux antérieurs sur l'alignement par segmentation binaire récursive ainsi que de travaux sur le clustering de documents. Nous évaluons les alignements produits sur des tâches de traduction automatique et montrons qu'il est possible d'atteindre des résultats du niveau de l'état de l'art, affichant des gains très conséquents allant jusqu'à plus de 4 points BLEU par rapport à nos travaux antérieurs, à l'aide une méthode très simple, indépendante de la taille du corpus à traiter, et produisant directement des alignements symétriques. En utilisant cette méthode en tant qu'extension à l'outil d'extraction de traductions Anymalign, nos expériences nous permettent de cerner certaines limitations de ce dernier et de définir des pistes pour son amélioration.

Dans cet article, nous nous intéressons à l'utilisation de la translittération arabe pour l'amélioration des résultats d'une approche linguistique d'alignement de mots simples et composés à partir de

corpus de textes parallèles français-arabe. Cette approche utilise, d'une part, un lexique bilingue et les caractéristiques linguistiques des entités nommées et des cognats pour l'alignement de mots simples, et d'autre part, les relations de dépendance syntaxique pour aligner les mots composés. Nous avons évalué l'aligneur de mots simples et composés intégrant la translittération arabe en utilisant deux procédés : une évaluation de la qualité d'alignement à l'aide d'un alignement de référence construit manuellement et une évaluation de l'impact de cet alignement sur la qualité de la traduction en faisant appel au système de traduction automatique statistique Moses. Les résultats obtenus montrent que la translittération améliore aussi bien la qualité de l'alignement que celle de la traduction.

Nous présentons dans cet article les améliorations apportées à la ressource « Les Verbes Français » afin de la rendre plus formelle et utilisable pour le traitement automatique des langues naturelles. Les informations syntaxiques et sémantiques ont été corrigées, restructurées, unifiées puis intégrées à la version XML de cette ressource, afin de pouvoir être utilisée par un système d'étiquetage de rôles sémantiques.

Cette étude vise à étudier les manifestations de la relation de méronymie dans une ressource lexicale générée automatiquement à partir d'un corpus de langue générale. La démarche que nous adoptons consiste à recueillir un jeu de couples de méronymes issus d'une ressource externe que nous croisons avec une base distributionnelle calculée à partir d'un corpus de textes encyclopédiques. Une annotation sémantique des mots qui entrent dans ces couples de méronymes montre que la prise en compte de la nature sémantique des mots composant les couples de méronymes permet de mettre au jour des inégalités au niveau du repérage de la relation par la méthode d'analyse distributionnelle.

Dans cet article, nous définissons un nouveau critère de cohésion thématique permettant de

pondérer les termes d'un lexique thématique en fonction de leur pertinence. Le critère s'inspire des approches Web as corpus pour accumuler des connaissances exogènes sur un lexique. Ces connaissances sont ensuite modélisées sous forme de graphe et un algorithme de marche aléatoire est appliqué pour attribuer un score à chaque terme. Après avoir étudié les performances et la stabilité du critère proposé, nous l'évaluons sur une tâche d'aide à la création de lexiques bilingues.

Ce travail présente des expériences initiales en validation de paraphrases en contexte. Les révisions de Wikipédia nous servent de domaine d'évaluation : pour un énoncé ayant connu une courte révision dans l'encyclopédie, nous disposons d'un ensemble de réécritures possibles, parmi lesquelles nous cherchons à identifier celles qui correspondent à des paraphrases valides. Nous abordons ce problème comme une tâche de classification fondée sur des informations issues du Web, et parvenons à améliorer la performance de plusieurs techniques simples de référence.

Cet article présente une méthode de simplification syntaxique de textes français. La simplification syntaxique a pour but de rendre des textes plus abordables en simplifiant les éléments qui posent problème à la lecture. La méthode mise en place à cette fin s'appuie tout d'abord sur une étude de corpus visant à étudier les phénomènes linguistiques impliqués dans la simplification de textes en français. Nous avons ainsi constitué un corpus parallèle à partir d'articles de Wikipédia et Vikidia, ce qui a permis d'établir une typologie de simplifications. Dans un second temps, nous avons implémenté un système qui opère des simplifications syntaxiques à partir de ces observations. Des règles de simplification ont été décrites afin de générer des phrases simplifiées. Un module sélectionne ensuite le meilleur ensemble de phrases. Enfin, nous avons mené une évaluation de notre système montrant qu'environ 80% des phrases générées sont correctes.

Dans cet article, nous proposons une étude comparative entre trois approches pour le résumé automatique de documents arabes. Ainsi, nous avons proposé trois méthodes pour l'extraction des

phrases les plus représentatives d'un document. La première méthode se base sur une approche symbolique, la deuxième repose sur une approche numérique et la troisième se base sur une approche hybride. Ces méthodes sont implémentées respectivement par le système ARSTResume, le système R.I.A et le système HybridResume. Nous présentons, par la suite, les résultats obtenus par les trois systèmes et nous procédons à une étude comparative entre les résultats obtenus afin de souligner les avantages et les limites de chaque méthode. Les résultats de l'évaluation ont montré que l?approche numérique est plus performante que l'approche symbolique au niveau des textes longs. Mais, l'intégration de ces deux approches en une approche hybride aboutit aux résultats les plus performants dans notre corpus de textes.

Cet article présente les résultats d'une analyse sémantique quantitative des unités lexicales spécifiques dans un corpus technique, relevant du domaine des machines-outils pour l'usinage des métaux. L'étude vise à vérifier si et dans quelle mesure les mots-clés du corpus technique sont monosémiques. A cet effet, nous procédons à une analyse statistique de régression simple, qui permet d'étudier la corrélation entre le rang de spécificité des mots-clés et leur rang de monosémie, mais qui soulève des problèmes statistiques et méthodologiques, notamment un biais de fréquence. Pour y remédier, nous adoptons une approche alternative pour le repérage des unités lexicales spécifiques, à savoir l'analyse des marqueurs lexicaux stables ou Stable Lexical Marker Analysis (SLMA). Nous discutons les résultats quantitatifs et statistiques de cette approche dans la perspective de la corrélation entre le rang de spécificité et le rang de monosémie.

Dans cet article, nous proposons une approche pour explorer des textes de taille importante en mettant en évidence des sous-parties cohérentes. Cette méthode d'exploration s'appuie sur une représentation en graphe du texte, en utilisant le modèle linguistique de Hoey pour sélectionner et apparier les phrases dans le graphe. Notre contribution porte sur l'utilisation de techniques de fouille de graphes sous contraintes pour extraire des sous-parties pertinentes du texte (c'est-à-dire des

collections de sous-réseaux phrastiques homogènes). Nous avons réalisé des expérimentations sur deux textes anglais de taille conséquente pour montrer l'intérêt de l'approche que nous proposons.

Cet article présente une étude détaillée de l'impact du type du corpus sur la tâche d'acquisition de paraphrases sous-phrastiques. Nos expériences sont menées sur deux langues et quatre types de corpus, et incluent une combinaison efficace de quatre systèmes d'acquisition de paraphrases. Nous obtenons une amélioration relative de plus de 27% en F-mesure par rapport au meilleur système, en anglais et en français, ainsi qu'une amélioration relative à notre combinaison de systèmes de 22% pour l'anglais et de 5% pour le français quand tous les types de corpus sont utilisés pour l'acquisition depuis le type de corpus le plus couramment disponible.

La qualité de l'annotation morpho-syntaxique d'un corpus est déterminante pour l'entraînement et l'évaluation de méthodes d'étiquetage. Cet article présente une série d'expériences que nous avons menée sur la détection et la correction automatique des erreurs du French Treebank. Deux méthodes sont utilisées. La première consiste à identifier les mots sans étiquette et leur attribuer celle d'une forme correspondante observée dans le corpus. La seconde méthode utilise les variations de n-gramme pour détecter et corriger les anomalies d'annotation. L'évaluation des corrections apportées au corpus est réalisée de manière extrinsèque en comparant les scores de performance de différentes méthodes d'étiquetage morpho-syntaxique en fonction du niveau de correction. Les résultats montrent une amélioration significative de la précision et indiquent que la qualité du corpus peut être sensiblement améliorée par l'application de méthodes de correction automatique des erreurs d'annotation.

Nous proposons d'annoter le French Treebank à l'aide de dépendances sémantiques dans le cadre de la DMRS en partant d'une annotation en dépendances syntaxiques de surface et en utilisant la réécriture modulaire de graphes. L'article présente un certain nombre d'avancées concernant le

calcul de réécriture utilisé : l'utilisation de règles pour faire le lien avec des lexiques, en particulier le lexique des verbes de Dicovalence, et l'introduction de filtres pour écarter à certaines étapes les annotations incohérentes. Il présente aussi des avancées dans le système de réécriture lui-même, qui a une plus large couverture (constructions causatives, verbes à montée, . . .) et dont l'ordre des modules a été étudié de façon plus systématique. Ce système a été expérimenté sur l'ensemble du French Treebank à l'aide du prototype GREW, qui implémente le calcul de réécriture utilisé.

Cet article présente les mécanismes de création d'un treebank hybride enrichissant le FTB à l'aide d'annotations dans le formalisme des Grammaires de Propriétés. Ce processus consiste à acquérir une grammaire GP à partir du treebank source et générer automatiquement les structures syntaxiques dans le formalisme cible en s'appuyant sur la spécification d'un schéma d'encodage adapté. Le résultat produit, en partant d'une version du FTB corrigée et modifiée en fonction de nos besoins, constitue une ressource ouvrant de nouvelles perspectives pour le traitement et la description du français.

Nous présentons dans cet article la méthodologie de constitution et les caractéristiques du corpus Sequoia, un corpus en français, syntaxiquement annoté d'après un schéma d'annotation très proche de celui du French Treebank (Abeillé et Barrier, 2004), et librement disponible, en constituants et en dépendances. Le corpus comporte des phrases de quatre origines : Europarl français, le journal l'Est Républicain, Wikipédia Fr et des documents de l'Agence Européenne du Médicament, pour un total de 3204 phrases et 69246 tokens. En outre, nous présentons une application de ce corpus : l'évaluation d'une technique d'adaptation d'analyseurs syntaxiques probabilistes à des domaines et/ou genres autres que ceux du corpus sur lequel ces analyseurs sont entraînés. Cette technique utilise des clusters de mots obtenus d'abord par regroupement morphologique à l'aide d'un lexique, puis par regroupement non supervisé, et permet une nette amélioration de l'analyse des domaines cibles (le corpus Sequoia), tout en préservant le même niveau de performance sur le domaine

source (le FTB), ce qui fournit un analyseur multi-domaines, à la différence d'autres techniques d'adaptation comme le self-training.

Cet article décrit une méthode permettant d'acquérir un lexique bilingue d'expressions polylexicales (EPLS) à partir d'un corpus parallèle français-anglais. Nous identifions dans un premier temps les EPLS dans chaque partie du corpus parallèle. Ensuite, nous proposons un algorithme d'alignement assurant la mise en correspondance bilingue d'EPLS. Pour mesurer l'apport du lexique construit, une évaluation basée sur la tâche de Traduction Automatique Statistique (TAS) est menée. Nous étudions les performances de trois stratégies dynamiques et d'une stratégie statique pour intégrer le lexique bilingue d'expressions polylexicales dans un système de TAS. Les expériences menées dans ce cadre montrent que ces unités améliorent significativement la qualité de traduction.

Dans cet article, nous nous focalisons sur la manière d'utiliser du clustering hiérarchique pour apprendre une grammaire AB à partir d'arbres de dérivation partiels. Nous décrirons brièvement les grammaires AB ainsi que les arbres de dérivation dont nous nous servons comme entrée pour l'algorithme, puis la manière dont nous extrayons les informations des corpus arborés pour l'étape de clustering. L'algorithme d'unification, dont le pivot est le cluster, sera décrit et les résultats analysés en détails.

Cet article présente une méthodologie permettant d'extraire et de décrire des locutions verbales vis-à-vis de leur comportement transformationnel. Plusieurs objectifs sont ciblés : 1) extraire automatiquement les expressions phraséologiques et en particulier les expressions figées, 2) décrire linguistiquement le comportement des phraséologismes 3) comparer les méthodes statistiques et notre approche et enfin 4) montrer l'importance de ces expressions dans un outil de classification de textes.

Les corpus multilingues sont extensivement exploités dans plusieurs branches du traitement automatique des langues. Cet article présente une vue d'ensemble des travaux en construction automatique de ces corpus. Nous traitons ce sujet en donnant premièrement un aperçu de différentes perceptions de la comparabilité. Nous examinons ensuite les principales approches de calcul de similarité, de construction et d'évaluation développées dans le domaine. Nous observons que Le calcul de la similarité textuelle se fait généralement sur la base de statistiques de corpus, de la structure de ressources ontologiques ou de la combinaison de ces deux approches. Dans un cadre multilingue avec l'utilisation d'un dictionnaire multilingue ou d'un traducteur automatique, de nombreux problèmes apparaissent. L'exploitation d'une ressource ontologique multilingue semble être une solution. En classification, la problématique de l'ajout de documents à la base initiale sans affecter la qualité des clusters demeure ouverte.

Dans cet article, nous présentons les étapes du développement de ressources pour l'entraînement et l'utilisation d'un nouvel outil de l'étiquetage morpho-syntaxique de la langue arabe. Nous avons mis en oeuvre un système basé sur l'étiqueteur stochastique TreeTagger, réputé pour son efficacité et la généricité de son architecture. Pour ce faire, nous avons commencé par la constitution de notre corpus de travail. Celui-ci nous a d'abord servi à réaliser l'étape de segmentation lexicale. Dans un second temps, ce corpus a permis d'effectuer l'entrainement de TreeTagger, grâce à un premier étiquetage réalisé avec l'étiqueteur ASVM 1.0, suivi d'une phase de correction manuelle. Nous détaillons ainsi les pré-traitements requis, et les différentes étapes de la phase d'apprentissage avec cet outil. Nous terminons par une évaluation sommaire des résultats, à la fois qualitative et quantitative. Cette évaluation, bien que réalisée sur un corpus de test de taille modeste, montre que nos premiers résultats sont encourageants.

Cet article décrit notre contribution sur la détection de polarité d'opinions en langue arabe par apprentissage supervisé. En effet le système proposé comprend trois phases: le pré-traitement du

corpus, l'extraction des caractéristiques et la classification. Pour la deuxième phase, nous utilisons vingt caractéristiques dont les principales sont l'émotivité, la réflexivité, l'adressage et la polarité. La phase de classification représente dans notre travail la combinaison des plusieurs classifieurs SVMs (Machine à Vecteur de Support) pour résoudre le problème multi classes. Nous avons donc analysés les deux stratégies de SVM multi classes qui sont : « un contre tous » et « un contre un » afin de comparer les résultats et améliorer la performance du système global.

Cet article présente les principales méthodes d'extraction automatique de termes-clés. La tâche d'extraction automatique de termes-clés consiste à analyser un document pour en extraire les expressions (phrasèmes) les plus représentatives de celui-ci. Les méthodes d'extraction automatique de termes-clés sont réparties en deux catégories : les méthodes supervisées et les méthodes non supervisées. Les méthodes supervisées réduisent la tâche d'extraction de termes-clés à une tâche de classification binaire (tous les phrasèmes sont classés parmi les termes-clés ou les non termes-clés). Cette classification est possible grâce à une phase préliminaire d'apprentissage, phase qui n'est pas requise par les méthodes non-supervisées. Ces dernières utilisent des caractéristiques (traits) extraites du document analysé (et parfois d'une collection de documents de références) pour vérifier des propriétés permettant d'identifier ses termes-clés.

Nous présentons dans cet article un système d'extraction de connaissances en arabe, fondé sur une analyse morpho-syntaxique profonde. Ce système reconnaît les mots simples, les expressions idiomatiques, les mots composés et les entités nommées. L'analyse identifie aussi les relations syntaxiques de dépendance et traite les formes passives et actives. L'extraction des connaissances est propre à l'application et utilise des règles d'extraction sémantiques qui s'appuient sur le résultat de l'analyse morpho-syntaxique. A ce niveau, le type de certaines entités nommées peut être révisé. L'extraction se base, dans nos expérimentations, sur une ontologie dans le domaine de la sécurité. Le RDF (Resource Description Framework) produit est ensuite traité pour regrouper les informations

qui concernent un même événement ou une même entité nommée. Les informations ainsi extraites peuvent alors aider à appréhender les informations contenues dans un ensemble de textes, alimenter une base de connaissances, ou bien servir à des outils de veille.

Nous présentons dans cet article les premiers résultats de nos travaux sur l'extraction de mots simples appartenant au lexique scientifique transdisciplinaire sur un corpus analysé morpho-syntaxiquement composé d'articles de recherche en sciences humaines et sociales. La ressource générée sera utilisée lors de l'indexation automatique de textes comme filtre d'exclusion afin d'isoler ce lexique de la terminologie. Nous comparons plusieurs méthodes d'extraction et montrons qu'un premier lexique de mots simples peut être dégagé et que la prise en compte des unités polylexicales ainsi que de la distribution seront nécessaires par la suite afin d'extraire l'ensemble de la phraséologie transdisciplinaire.

Cette recherche porte sur le chiasme de mots : figure de style jouant sur la réversion (ex. « Bonnet blanc, blanc bonnet »). Elle place le chiasme dans la problématique de sa reconnaissance automatique : qu'est-ce qui le définit et comment un ordinateur peut le trouver ? Nous apportons une description formelle du phénomène. Puis nous procédons à la constitution d'une liste d'exemples contextualisés qui nous sert au test des hypothèses. Nous montrons ainsi que l'ajout de contraintes formelles (contrôle de la ponctuation et omission des mots vides) pénalise très peu le rappel et augmente significativement la précision de la détection. Nous montrons aussi que la lemmatisation occasionne peu d'erreurs pour le travail d'extraction mais qu'il n'en est pas de même pour la racinisation. Enfin nous mettons en évidence que l'utilisation d'un thésaurus apporte quelques résultats pertinents.

Dans cet article nous nous intéressons au choix d'un formalisme de représentation des connaissances qui nous permette de représenter, manipuler, interroger et raisonner sur des

connaissances linguistiques du Dictionnaire Explicatif et Combinatoire (DEC) de la Théorie Sens-Texte. Nous montrons que ni les formalismes du web sémantique ni le formalisme des Graphes conceptuels n'est adapté pour cela, et justifions l'introduction d'un nouveau formalisme dit des Graphes d'Unités. Nous introduisons la hiérarchie des Types d'Unités au coeur du formalisme, et présentons les Graphes d'Unités ainsi que la manière dont on peut les utiliser pour représenter certains aspects du DEC.

Ce travail présente une nouvelle approche pour injecter des dépendances profondes (sujet des verbes à contrôle, partage du sujet en cas d'ellipses, ...) dans un corpus arboré présentant un schéma d'annotation surfacique et projectif. Nous nous appuyons sur un système de réécriture de graphes utilisant des techniques de programmation par contraintes pour produire des règles génériques qui s'appliquent aux phrases du corpus. Par ailleurs, nous testons la généricité des règles en utilisant des sorties de trois analyseurs syntaxiques différents, afin d'évaluer la dégradation exacte de l'application des règles sur des analyses syntaxiques prédites.

La désambiguïsation lexicale, le processus qui consiste à automatiquement identifier le ou les sens possible d'un mot polysémique dans un contexte donné, est une tâche fondamentale pour le Traitement Automatique des Langues (TAL). Le développement et l'amélioration des techniques de désambiguïsation lexicale ouvrent de nombreuses perspectives prometteuses pour le TAL. En effet, cela pourrait conduire à un changement paradigmatique en permettant de réaliser un premier pas vers la compréhension des langues naturelles. En raison du manque de ressources langagières, il est parfois difficile d'appliquer des techniques de désambiguïsation à des langues peu dotées. C'est pourquoi, nous nous intéressons ici, à enquêter sur comment avoir un début de recherche sur la désambiguïsation lexicale pour les langues peu dotées, en particulier en exploitant des techniques d'induction des sens de mots, ainsi que quelques suggestions de pistes intéressantes à explorer.

Ces derniers temps, vu la situation préoccupante du monde arabe, les dialectes arabes et notamment le dialecte tunisien est devenu de plus en plus utilisé dans les interviews, les journaux télévisés et les émissions de débats. Cependant, cette situation présente des conséquences négatives importantes pour le Traitement Automatique du Langage Naturel (TALN): depuis que les dialectes parlés ne sont pas officiellement écrits et n'ont pas d'orthographe standard, il est très coûteux d'obtenir des corpus adéquats à utiliser pour des outils de TALN. Par conséquent, il n'existe pas des corpus parallèles entre l'Arabe Standard Moderne(ASM) et le Dialecte Tunisien (DT). Dans ce travail, nous proposons une méthode pour la création d'un lexique bilingue ASM?DT et un processus pour la génération automatique de corpus dialectaux. Ces ressources vont servir à la construction d'un modèle de langage pour les journaux télévisés tunisiens, afin de l'intégrer dans un Système de Reconnaissance Automatique de Parole (SRAP).

Les utilisateurs d'un système de recherche d'information mettent en oeuvre des comportements de recherche complexes tels que la reformulation de requête et la recherche multitâche afin de satisfaire leurs besoins d'information. Ces comportements de recherche peuvent être observés à travers des journaux de requêtes, et constituent des indices permettant une meilleure compréhension des besoins des utilisateurs. Dans cette perspective, il est nécessaire de regrouper au sein d'une même session de recherche les requêtes reliées à un même besoin d'information. Nous proposons une méthode de détection automatique des sessions exploitant la collection de documents WIKIPÉDIA, basée sur la similarité des résultats renvoyés par l'interrogation de cette collection afin d'évaluer la similarité entre les requêtes. Cette méthode obtient de meilleures performances que les approches temporelle et lexicale traditionnellement employées pour la détection de sessions séquentielles, et peut être appliquée à la détection de sessions imbriquées. Ces expérimentations ont été réalisées sur des données provenant du portail OpenEdition.

Cet article présente une approche mixte, morpho-syntaxique et statistique, pour la reconnaissance

d'entités nommées en langue chinoise dans un système d'extraction automatique d'information. Le processus se divise principalement en trois étapes : la première génère des noms propres potentiels à l'aide de règles morphologiques ; la deuxième utilise un modèle de langue afin de sélectionner le meilleur résultat ; la troisième effectue la reconnaissance d'entités nommées grâce à une analyse syntaxique locale. Cette dernière permet une reconnaissance automatique d'entités nommées plus pertinente et plus complète.

La charte Ethique & Big Data a été conçue à l'initiative de l'ATALA, de l'AFCP, de l'APROGED et de CAP DIGITAL, au sein d'un groupe de travail mixte réunissant d'autres partenaires académiques et industriels (tels que le CERSA-CNRS, Digital Ethics, Eptica-Lingway, le cabinet Itéanu ou ELRA/ELDA). Elle se donne comme objectif de fournir des garanties concernant la traçabilité des données (notamment des ressources langagières), leur qualité et leur impact sur l'emploi. Cette charte a été adoptée par Cap Digital (co-rédacteur). Nous avons également proposé à la DGLFLF et à l'ANR de l'utiliser. Elle est aujourd'hui disponible sous forme de wiki, de fichier pdf et il en existe une version en anglais. La charte est décrite en détails dans (Couillault et Fort, 2013).

La recherche scientifique est un processus incrémental. La première étape à effectuer avant de débuter des travaux consiste à réaliser un état de l'art des méthodes existantes. La communauté francophone du Traitement Automatique de la Langue (TAL) produit de nombreuses publications scientifiques qui sont malheureusement dispersées sur différents sites et pour lesquelles aucune méta-donnée n'est disponible. Cet article présente la construction de TALN Archives, une archive numérique francophone des articles de recherche en TAL dont le but est d'offrir un accès simplifié aux différents travaux effectués dans notre domaine. Nous présentons également une analyse du réseau de collaboration construit à partir des méta-données que nous avons extraites et dévoilons l'identité du Kevin Bacon de TALN Archives, i.e. l'auteur le plus central dans le réseau de collaboration.

En présence de corpus comparables bilingues, nous sommes confrontés à des données qu'il est naturel de plonger dans deux espaces de représentation linguistique distincts, chacun éventuellement muni d'une mesure quantifiable de similarité (ou d'une distance). Dès lors que ces données bilingues sont comparables au sens d'une mesure de comparabilité également calculable (Li et Gaussier, 2010), nous pouvons établir une connexion entre ces deux espaces de représentation linguistique en exploitant une carte d'association pondérée ("mapping") appréhendée sous la forme d'un graphe bi-directionnel dit de comparabilité. Nous abordons dans cet article les conséquences conceptuelles et pratique d'une telle connexion similarité-comparabilité en développant un algorithme (Hit-ComSim) basé sur sur le principe de similarité induite par la topologie du graphe de comparabilité. Nous essayons de qualifier qualitativement l'intérêt de cet algorithme en considérant quelques expériences préliminaires de clustering de documents comparables bilingues (Français/Anglais) collectés sur des flux RSS.

ProLMF est la version LMF de la base lexicale multi-lingue de noms propres Prolexbase. Disponible librement sur le site du CNRTL, la version 1.2 a été largement améliorée et augmentée par de nouvelles entrées en français, complétées par des expansions contextuelles, et par de petits lexiques en une huitaine de langues.

Récemment, le paradigme du décodage guidé a montré un fort potentiel dans le cadre de la reconnaissance automatique de la parole. Le principe est de guider le processus de décodage via l'utilisation de transcriptions auxiliaires. Ce paradigme appliqué à la traduction automatique permet d'envisager de nombreuses applications telles que la combinaison de systèmes, la traduction multi-sources etc. Cet article présente une approche préliminaire de l'application de ce paradigme à la traduction automatique (TA). Nous proposons d'enrichir le modèle log-linéaire d'un système primaire de TA avec des mesures de distance relatives à des systèmes de TA auxiliaires. Les

premiers résultats obtenus sur la tâche de traduction Français/Anglais issue de la campagne d'évaluation WMT 2011 montrent le potentiel du décodage guidé.

Cet article s'intéresse à la traduction automatique statistique des forums, dans le cadre du projet européen ACCEPT (« Automated Community Content Editing Portal »). Nous montrons qu'il est possible d'écrire des règles de préédition peu coûteuses sur le plan des ressources linguistiques et applicables sans trop d'effort avec un impact très significatif sur la traduction automatique (TA) statistique, sans avoir à modifier le système de TA. Nous décrivons la méthodologie proposée pour écrire les règles de préédition et les évaluer, ainsi que les résultats obtenus par type de règles.

Les avatars signeurs en Langue des Signes Française (LSF) sont de plus en plus utilisés en tant qu'interface de communication à destination de la communauté sourde. L'un des critères d'acceptation de ces avatars est l'aspect naturel et réaliste des gestes produits. Par conséquent, des méthodes de synthèse de gestes ont été élaborées à l'aide de corpus de mouvements capturés et annotés provenant d'un signeur réel. Néanmoins, l'enrichissement d'un tel corpus, en faisant fi des séances de captures supplémentaires, demeure une problématique certaine. De plus, l'application automatique d'opérations sur ces mouvements (e.g. concaténation, mélange, etc.) ne garantit pas la consistance sémantique du geste résultant. Une alternative est d'insérer l'opérateur humain dans la boucle de construction des énoncés en LSF. Dans cette optique, cet article propose un premier système interactif d'édition de gestes en LSF, basé "données capturées" et dédié aux avatars signeurs.

Cet article présente la réalisation d'ANCOR, qui constitue par son envergure (453 000 mots) le premier corpus francophone annoté en anaphores et co-références permettant le développement d'approches centrées sur les données pour la résolution des anaphores et autres traitements de la co-référence. L'annotation a été réalisée sur trois corpus de parole conversationnelle (Accueil_UBS,

OTG et ESLO) qui le destinent plus particulièrement au traitement du langage parlé. En l'absence d'équivalent pour le langage écrit, il est toutefois susceptible d'intéresser l'ensemble de la communauté TAL. Par ailleurs, le schéma d'annotation retenu est suffisamment riche pour permettre des études en linguistique de corpus. Le corpus sera diffusé librement à la mi-2013 sous licence Creative Commons BY-NC-SA. Cet article se concentre sur sa mise en oeuvre et décrit brièvement quelques résultats obtenus sur la partie déjà annotée de la ressource.

La composition est un phénomène fréquent dans plusieurs langues, surtout dans des langues ayant une morphologie riche. Le traitement des mots composés est un défi pour les systèmes de TAL car pour la plupart, ils ne sont pas présents dans les lexiques. Dans cet article, nous présentons une méthode de segmentation des composés qui combine des caractéristiques indépendantes de la langue (mesure de similarité, données du corpus) avec des règles de transformation sur les frontières des composants spécifiques à une langue. Nos expériences de segmentation de termes composés allemands et russes montrent une exactitude jusqu'à 95 % pour l'allemand et jusqu'à 91 % pour le russe. Nous constatons que l'utilisation de corpus spécialisés relevant du même domaine que les composés améliore la qualité de segmentation.

La gestion des terminologies pose encore des problèmes, en particulier pour des constructions complexes comme les acronymes. Dans cet article, nous proposons une solution en reliant plusieurs termes différents à un seul référent via les notions de pivot et de prolexème. Ces notions permettent par exemple de faire le lien entre plusieurs termes qui désignent un même et unique référent : Nations Unies, ONU, Organisation des Nations Unies et onusien. Il existe Jibiki, une plate-forme générique de gestion de bases lexicales permettant de gérer n'importe quel type de structure (macro et microstructure). Nous avons implémenté une nouvelle macrostructure de ProAxie dans la plate-forme Jibiki pour réaliser la gestion des acronymes.

Notre article présente expérimentations portant sur la classification supervisée de variétés nationales de l'espagnol. Outre les approches classiques, basées sur l'utilisation de n-grammes de caractères ou de mots, nous avons testé des modèles calculés selon des traits morpho-syntaxiques, l'objectif étant de vérifier dans quelle mesure il est possible de parvenir à une classification automatique des variétés d'une langue en s'appuyant uniquement sur des descripteurs grammaticaux. Les calculs ont été effectués sur la base d'un corpus de textes journalistiques de quatre pays hispanophones (Espagne, Argentine, Mexique et Pérou).

La majeure partie des travaux en fouille d'opinion et en analyse de sentiment concerne le classement des opinions majoritaires. Les méthodes d'apprentissage supervisé à base de n-grammes sont souvent employées. Elles ont l'inconvénient d'avoir un biais en faveur des opinions majoritaires si on les utilise de manière classique. En fait la présence d'un terme particulier, fortement associé à la cible de l'opinion dans un document peut parfois suffire à faire basculer le classement de ce document dans la classe de ceux qui expriment une opinion majoritaire sur la cible. C'est un phénomène positif pour l'exactitude globale du classifieur, mais les documents exprimant des opinions minoritaires sont souvent mal classés. Ce point est un problème dans le cas où l'on s'intéresse à la détection des signaux faibles (détection de rumeur) ou pour l'anticipation de renversement de tendance. Nous proposons dans cet article d'améliorer la classification des opinions minoritaires en prenant en compte les Entités Nommées dans le calcul de pondération destiné à corriger le biais en faveur des opinions majoritaires.

Au cours des deux dernières décennies des psychologues et des linguistes informaticiens ont essayé de modéliser l'accès lexical en construisant des simulations ou des ressources. Cependant, parmi ces chercheurs, pratiquement personne n'a vraiment cherché à améliorer la navigation dans des 'dictionnaires électroniques destinés aux producteurs de langue'. Pourtant, beaucoup de travaux ont été consacrés à l'étude du phénomène du mot sur le bout de la langue et à la

construction de réseaux lexicaux. Par ailleurs, vu les progrès réalisés en neurosciences et dans le domaine des réseaux complexes, on pourrait être tenté de construire un simulacre du dictionnaire mental, ou, à défaut une ressource destinée aux producteurs de langue (écrivains, conférenciers). Nous sommes restreints en construisant un réseau de co-occurrences à partir des résumés de Wikipedia, le but étant de vérifier jusqu'où l'on pouvait pousser une telle ressource pour trouver un mot, sachant que la ressource ne contient pas de liens sémantiques, car le réseau est construit de manière automatique et à partir de textes non-annotés.

Nous nous intéressons ici à la construction semi-automatique de grammaires computationnelles et à leur utilisation pour l'analyse syntaxique. Nous considérons des grammaires lexicalisées dont les structures élémentaires sont des arbres, sous-spécifiés ou pas. Nous présentons un algorithme qui vise à prévoir l'ensemble des arbres élémentaires attachés aux mots qui peuvent s'intercaler entre deux mots donnés d'une phrase, dont on sait que les arbres élémentaires associées sont des compagnons, c'est-à-dire qu'ils interagiront nécessairement dans la composition syntaxique de la phrase.

La série de campagnes d'évaluation CoNLL-2011/2012 a permis de comparer diverses propositions d'architectures de systèmes de détection de co-références. Cet article décrit le système de résolution de co-référence Poly-co développé dans le cadre de la campagne d'évaluation CoNLL-2011 et évalue son potentiel d'amélioration en introduisant des propriétés sémantiques dans son modèle de détection. Notre système s'appuie sur un classifieur perceptron multi-couches. Nous décrivons les heuristiques utilisées pour la sélection des paires de mentions candidates, ainsi que l'approche de sélection des traits caractéristiques que nous avons utilisée lors de la campagne CoNLL-2011. Nous introduisons ensuite un trait sémantique complémentaire et évaluons son influence sur les performances du système.

Dans cet article, nous nous intéressons au pré-traitement de la langue arabe comme langue source à des fins de traduction automatique statistique. Nous présentons une étude sur la traduction automatique statistique basée sur les syntagmes, pour la paire de langues arabe-français utilisant le décodeur Moses ainsi que d'autres outils de base. Les propriétés morphologiques et syntaxiques de la langue arabe sont complexes, ce qui rend cette langue difficile à maîtriser dans le domaine du TALN. Aussi, les performances d'un système de traduction statistique dépendent considérablement de la quantité et de la qualité des corpus d'apprentissage. Dans cette étude, nous montrons qu'un pré-traitement basé sur les mots de la langue source (arabe) et l'introduction de quelques règles linguistiques par rapport à la syntaxe de la langue cible (français), permet d'obtenir des améliorations du score BLEU. Cette amélioration est réalisée sans augmenter la quantité des corpus d'apprentissage.

Nous proposons dans cet article une méthodologie, qui s'inspire du développement agile et qui permettrait d'assister la préparation d'une campagne d'annotation. Le principe consiste à formaliser au maximum les instructions contenues dans le guide d'annotation afin de vérifier automatiquement si le corpus en construction est cohérent avec le guide en cours d'écriture. Pour exprimer la partie formelle du guide, nous utilisons la réécriture de graphes, qui permet de décrire par des motifs les constructions définies. Cette formalisation permet de repérer les constructions prévues par le guide et, par contraste, celles qui ne sont pas cohérentes avec le guide. En cas d'incohérence, un expert peut soit corriger l'annotation, soit mettre à jour le guide et relancer le processus.

Les titres de cartes topographiques personnalisées composent un corpus spécifique caractérisé par des variations orthographiques et un nombre élevé de désignations de lieux. L'article présente le repérage des toponymes dans ces titres. Ce repérage est fondé sur l'utilisation de BDNyme, la base de données de toponymes géoréférencés de l'IGN, et sur une analyse de surface à l'aide de patrons. La méthode proposée élargit la définition du toponyme pour tenir compte de la nature du

corpus et des données qu'il contient. Elle se décompose en trois étapes successives qui tirent parti du contexte extralinguistique de géoréférencement des toponymes et du contexte linguistique. Une quatrième étape qui ne retient pas le géoréférencement est aussi étudiée. Le balisage et le typage des toponymes permettent de mettre en avant d'une part la diversité des désignations de lieux et d'autre part leurs variations d'écriture. La méthode est évaluée (rappel, précision, F-mesure) et les erreurs analysées.

Le défi i2b2/VA 2012 était dédié à la détection de relations temporelles entre événements et expressions temporelles dans des comptes rendus hospitaliers en anglais. Les situations considérées étaient beaucoup plus variées que dans les défis TempEval. Nous avons donc axé notre travail sur un examen systématique de 57 situations différentes et de leur importance dans le corpus d'apprentissage en utilisant un oracle, et avons déterminé empiriquement le classifieur qui se comportait le mieux dans chaque situation, atteignant ainsi une F-mesure globale de 0,623.

Cet article décrit l'utilisation de la technique de similarité de second ordre pour l'identification de textes semblables au sein d'une base de rapports d'incidents aéronautiques mélangeant les langues française et anglaise. L'objectif du système est, pour un document donné, de retrouver des documents au contenu similaire quelle que soit leur langue. Nous utilisons un corpus bilingue aligné de rapports d'accidents aéronautiques pour construire des paires de pivots et indexons les documents avec des vecteurs de similarités, tels que chaque coordonnée correspond au score de similarité entre un document dans une langue donnée et la partie du pivot de la même langue. Nous évaluons les performances du système sur un volumineux corpus de rapports d'incidents aéronautiques pour lesquels nous disposons de traductions. Les résultats sont prometteurs et valident la technique.

La catégorisation de textes nécessite généralement un investissement important en amont, avec

une adaptation de domaine. L'approche que nous proposons ici permet d'associer finement à un texte tout-venant écrit dans une langue donnée, un graphe de catégories de la Wikipédia dans cette langue. L'utilisation de l'index inter-langues de l'encyclopédie en ligne permet de plus d'obtenir un sous-ensemble de ce graphe dans la plupart des autres langues.

Cet article présente une méthode hybride d'enrichissement d'un lexique de noms propres à partir de la base encyclopédique en ligne Wikipedia. Une des particularités de cette recherche est de viser l'enrichissement d'une ressource existante (Prolexbase) très contrôlée décrivant finement les noms propres. A la différence d'autres travaux destinés à la reconnaissance des entités nommées, notre objectif est donc de réaliser un enrichissement automatique de qualité. Notre approche repose sur l'utilisation en pipe-line de règles déterministes basées sur certaines informations DBpedia et d'une catégorisation supervisée à base de classifieur SVM. Nos résultats montrent qu'il est ainsi possible d'enrichir un lexique de noms propres avec une très bonne précision.

Nous présentons ici les premiers travaux concernant l'établissement d'une passerelle bi-directionnelle entre d'une, part les schémas d'annotation syntaxique en dépendances qui ont été définis pour convertir les annotations du French Treebank en arbres de dépendances de surface pour l'analyseur syntaxique Bonsai, et d'autre part le formalisme d'annotation PASSAGE développé initialement pour servir de support à des campagnes d'évaluation ouvertes en mode objectif quantitatif boîte-noire pour l'analyse syntaxique du français.

La traduction des pronoms est l'un des problèmes actuels majeurs en traduction automatique. Étant donné que les pronoms ne transmettent pas assez de contenu sémantique en eux-mêmes, leur traitement automatique implique la résolution des anaphores. La recherche en résolution des anaphores s'intéresse à établir le lien entre les entités sans contenu lexical (potentiellement des syntagmes nominaux et pronoms) et leurs référents dans le texte. Dans cet article, nous mettons en

oeuvre un premier prototype d'une méthode inspirée de la théorie du liage chomskyenne pour l'interprétation des pronoms dans le but d'améliorer la traduction des pronoms personnels entre l'espagnol et le français.

Nous évaluons l'utilité de trois lexiques bilingues dans un cadre de recherche interlingue français vers anglais sur le corpus CLEF. Le premier correspond à un dictionnaire qui couvre le corpus, alors que les deux autres ont été construits automatiquement à partir des sous-ensembles français et anglais de CLEF, en les considérant comme des corpus comparables. L'un contient des mots simples, alors que le deuxième ne contient que des termes complexes. Les lexiques sont intégrés dans des interfaces différentes dont les performances de recherche interlingue sont évaluées par 5 utilisateurs sur 15 thèmes de recherche CLEF. Les meilleurs résultats sont obtenus en intégrant le lexique de mots simples généré à partir des corpus comparables dans une interface proposant les cinq « meilleures » traductions pour chaque mot de la requête.

Cet article présente deux méthodes permettant de corriger des réclamations contenant des erreurs rédactionnelles, en s'appuyant sur le graphe des voisins orthographiques et contextuels. Ce graphe est constitué des formes ou mots trouvés dans un corpus d'apprentissage. Un lien entre deux formes traduit le fait que les deux formes se « ressemblent » et partagent des contextes similaires. La première méthode est semi-automatique et consiste à produire un dictionnaire de substitution à partir de ce graphe. La seconde méthode, plus ambitieuse, est entièrement automatisée. Elle s'appuie sur les contextes pour déterminer à quel mot correspond telle forme abrégée ou erronée. Les résultats ainsi obtenus permettent d'améliorer le processus déjà existant de constitution d'un dictionnaire de substitution mis en place au sein d'EDF.

Le marché d'offres d'emploi et des candidatures sur Internet a connu, ces derniers temps, une croissance exponentielle. Ceci implique des volumes d'information (majoritairement sous la forme

de textes libres) intraitables manuellement. Les CV sont dans des formats très divers : .pdf, .doc, .dvi, .ps, etc., ce qui peut provoquer des erreurs lors de la conversion en texte plein. Nous proposons SegCV, un système qui a pour but l'analyse automatique des CV des candidats. Dans cet article, nous présentons des algorithmes reposant sur une analyse de surface, afin de segmenter les CV de manière précise. Nous avons évalué la segmentation automatique selon des corpus de référence que nous avons constitués. Les expériences préliminaires réalisées sur une grande collection de CV en français avec correction du bruit montrent de bons résultats en précision, rappel et F-Score.

Les recherches présentées sont directement liées aux travaux menés pour résoudre les problèmes de catégorisation automatique de texte. Les mots porteurs d'opinions jouent un rôle important pour déterminer l'orientation du message. Mais il est essentiel de pouvoir identifier les cibles auxquelles ils se rapportent pour en contextualiser la portée. L'analyse peut également être menée dans l'autre sens, on cherchant dans le contexte d'une cible détectée les termes polarisés. Une première étape d'apprentissage depuis des données permet d'obtenir automatiquement les marqueurs de polarité les plus importants. A partir de cette base, nous cherchons les cibles qui apparaissent le plus fréquemment à proximité de ces marqueurs d'opinions. Ensuite, nous construisons un ensemble de couples (marqueur de polarité, cible) pour montrer qu'en s'appuyant sur ces couples, on arrive à expliquer plus finement les prises de positions tout en maintenant (voire améliorant) le niveau de performance du classifieur.

Avec le développement de la post-édition, de plus en plus de corpus contenant des corrections de traductions sont disponibles. Ce travail présente un corpus de corrections d'erreurs de traduction collecté dans le cadre du projet ANR/TRACE et illustre les différents types d'analyses auxquels il peut servir. Nous nous intéresserons notamment à la détection des erreurs fréquentes et à l'analyse de la variabilité des post-éditions.

Cet article présente une méthode automatique d'évaluation du contenu des résumés automatiques. La méthode proposée est basée sur une combinaison de caractéristiques englobant des scores de contenu et d'autres de complexité textuelle et ce en s'appuyant sur une technique d'apprentissage, à savoir la régression linéaire. L'objectif de cette combinaison consiste à prédire le score manuel PYRAMID à partir des caractéristiques utilisées. Afin d'évaluer la méthode présentée, nous nous sommes intéressés à deux niveaux de granularité d'évaluation : la première est qualifiée de Micro-évaluation et propose l'évaluation de chaque résumé, alors que la deuxième est une Macro-évaluation et s'applique au niveau de chaque système.

Dans cet article, nous nous intéressons à la segmentation thématique d'émissions télévisées exploitant la cohésion lexicale. Le but est d'étudier une approche générique, reposant uniquement sur la transcription automatique sans aucune information externe ni aucune information structurelle sur le contenu traité. L'étude porte plus particulièrement sur le mécanisme de pondération des mots utilisés lors du calcul de la cohésion lexicale. Les poids TF-IDF sont estimés à partir du contenu lui-même, qui est considéré comme une collection de documents mono-thème. Nous proposons une approche itérative, intégrée à un algorithme de segmentation, visant à raffiner la partition du contenu en documents pour l'estimation de la pondération. La segmentation obtenue à une itération donnée fournit un ensemble de documents à partir desquels les poids TF-IDF sont ré-estimés pour la prochaine itération. Des expériences menées sur un corpus couvrant différents formats des journaux télévisés issus de 8 chaînes françaises montrent une amélioration du processus global de segmentation.

Nous présentons PatternSim, une nouvelle mesure de similarité sémantique qui repose d'une part sur des patrons lexico-syntaxiques appliqués à de très vastes corpus et d'autre part sur une formule de ré-ordonnancement des candidats extraits. Le système, initialement développé pour l'anglais, a

été adapté au français. Nous rendons compte de cette adaptation, nous en proposons une évaluation et décrivons l'usage de ce nouveau modèle dans la plateforme de consultation en ligne Serelex.

Nous décrivons l'organisation et l'état courant de l'analyseur morphologique de l'allemand AMALD de grande taille couvrant (près de 103000 lemmes et 500000 formes fléchies simples, en croissance) développé dans le cadre du projet ANR-Émergence Traouiero. C'est le premier lemmatiseur de l'allemand capable de traiter non seulement les mots simples et les mots composés, mais aussi les verbes à particules séparables quand elles sont séparées, même par un grand nombre de mots (ex : Hier schlagen wir eine neue Methode für die morphologische Analyse vor).

Ce travail présente un corpus en français dédié à l'analyse de sentiment. Nous y décrivons la construction et l'organisation du corpus. Nous présentons ensuite les résultats de l'application de techniques d'apprentissage automatique pour la tâche de classification d'opinion (positive ou négative) véhiculée par un texte. Deux techniques sont utilisées : la régression logistique et la classification basée sur des Support Vector Machines (SVM). Nous mentionnons également l'intérêt d'appliquer une sélection de variables avant la classification (par régularisation par elastic net).

Le traitement des collocations en analyse et en traduction est depuis de nombreuses années au centre de nos intérêts de recherche. L'analyseur Fips a été récemment enrichi d'un module de résolution d'anaphores. Dans cet article nous décrivons comment la résolution d'anaphores a été appliquée à l'identification des collocations et comment cela permet à l'analyseur de repérer une collocation même si un de ses termes a été pronominalisé. Nous décrivons aussi la méthodologie de l'évaluation, notamment la préparation des données pour le calcul du rappel. Dans la tâche d'identification des collocations pronominalisées, Fips montre des résultats très encourageants : la précision mesurée est de 98% alors que le rappel est proche de 50%. Dans cette évaluation nous

nous intéressons aux collocations de type verbe-objet direct en conjonction avec les pronoms anaphoriques à la 3e personne. Le corpus utilisé est un corpus anglais d'environ dix millions de mots.

En s'appuyant sur une expérience d'enrichissement terminologique, cet article montre comment assister le travail d'acquisition terminologique et surmonter concrètement les deux difficultés qu'il présente : la masse de candidats-termes à considérer et la subjectivité des jugements terminologiques qui varient notamment en fonction du type de terminologie à produire. Nous proposons des stratégies simples pour filtrer a priori une partie du bruit des résultats des extracteurs et rendre ainsi la validation praticable pour des terminologues et nous démontrons leur efficacité sur un échantillon de candidats-termes proposés à la validation de deux spécialistes du domaine. Nous montrons également qu'en appliquant à une campagne de validation terminologique les mêmes principes méthodologiques que pour une campagne d'annotation, on peut contrôler la qualité des jugements de validation posés et de la terminologie qui en résulte.

DAnIEL est un système multi-lingue de veille épidémiologique. DAnIEL permet de traiter un grand nombre de langues à faible coût grâce à une approche parcimonieuse en ressources.

Le lexique multi-lingue basé sur une hiérarchie sémantique universelle fait partie du modèle linguistique Compreno destiné à plusieurs applications du TALN, y compris la traduction automatique et l'analyse sémantique et syntaxique. La ressource est propriétaire et n'est pas librement disponible.

Avec la digitalisation massive de documents apparaît la nécessité de disposer de systèmes de recherche capables de s'adapter aux habitudes de recherche des utilisateurs et de leur permettre d'accéder à l'information rapidement et efficacement. INBENTA a ainsi créé un moteur de recherche

intelligent appellé Inbenta semantic Search Engine (ISSE). Les deux tâches principales de l'ISSE sont d'analyser les questions des utilisateurs et de trouver la réponse appropriée à la requête en effectuant une recherche dans une base de connaissances. Pour cela, la solution logicielle d'INBENTA se base sur la Théorie Sens-Texte qui se concentre sur le lexique et la sémantique.

Nous décrivons notre prototype d'analyse automatique d'opinion. Celui-ci est basé sur un moteur d'analyse linguistique. Il permet de détecter finement les segments de texte porteurs d'opinions, de les extraire, et de leur attribuer une note selon la polarité qu'ils expriment. Nous présentons enfin les différentes perspectives que nous envisageons pour ce prototype.

Synapse Développement souhaite échanger avec les conférenciers autour des technologies qu'elle commercialise : correction de textes et analyse sémantique. Plusieurs produits et démonstrateurs seront présentés, notre but étant d'instaurer un dialogue et de confronter notre approche du TAL, à base de méthodes symboliques et statistiques influencées par des contraintes de production, et celles utilisées par les chercheurs, industriels ou passionnés qui viendront à notre rencontre.

Cet article décrit une interface graphique de visualisation de chronologies événementielles construites automatiquement à partir de requêtes thématiques en utilisant un corpus de dépêches fourni par l'Agence France Pressse (AFP). Cette interface permet également la validation des chronologies par des journalistes qui peuvent ainsi les éditer et les modifier.

CasSys est un système de création et de mise en oeuvre de cascades de transducteurs intégré à la plateforme Unitex. Nous présentons dans cette démonstration la nouvelle version implantée fin 2012. En particulier ont été ajoutées une interface plus conviviale et la possibilité d'itérer un même transducteur jusqu'à ce qu'il n'ait plus d'influence sur le texte. Un premier exemple concernera le traitement de texte avec une gestion complexe de balises XML et un deuxième présentera la

cascade CasEN de reconnaissance des entités nommées.

Une passerelle interactive d'accès multi-lingue (iMAG) dédiée à un site Web S (iMAG--?S) est un bon outil pour rendre S accessible dans beaucoup de langues, immédiatement et sans responsabilité éditoriale. Les visiteurs de S ainsi que des post--?éditeurs et des modérateurs payés ou non contribuent à l'amélioration continue et incrémentale des segments textuels les plus importants, et éventuellement de tous. Dans cette approche, les pré--?traductions sont produites par un ou plusieurs systèmes de Traduction Automatique (TA) gratuits. Il y a deux effets de bord intéressants, obtenables sans coût additionnel : les iMAGs peuvent être utilisées pour produire des corpus parallèles de haute qualité, et pour mettre en place une évaluation permanente et finalisée de multiples systèmes de TA.

Nous présentons un système basé sur les technologies du Web Sémantique pour la gestion, le développement et l'exploitation de données lexicales en réseau (Lexical Linked Data, LLD).

La plateforme ScienQuest fut initialement créée pour l'étude linguistique du positionnement et du raisonnement dans le corpus Scientext. Cette démonstration présente les modifications apportées à cette plateforme, pour en faire une base phraséologique adaptée à l'aide à la rédaction en langue seconde. Cette adaptation est utilisée dans le cadre de deux expérimentations en cours : l'aide à la rédaction en anglais pour les scientifiques, et l'aide à la rédaction académique en français pour les apprenants.

Le démonstrateur Apopsis permet de délimiter et de catégoriser les opinions émises sur les tweets en temps réel pour un sujet choisi par l'utilisateur au travers d'une interface web.

Les conversations téléphoniques qui contiennent du sentiment négatif sont particulièrement

intéressantes pour les centres d'appels, aussi bien pour évaluer la perception d'un produit par les clients que pour améliorer la formation des télé-conseillers. Néanmoins, ces conversations sont peu nombreuses et difficiles à trouver dans la masse d'enregistrements. Nous présentons un module d'analyse des sentiments qui permet de visualiser le déroulement émotionnel des conversations. Il se greffe sur un moteur de recherche, ce qui permet de trouver rapidement les conversations problématiques grâce à l'ordonnancement par score de négativité.

TermSuite est outil libre multi-lingue réalisant une extraction terminologique monolingue et une extraction terminologique bilingue à partir de corpus comparables.

Si les techniques statistiques pour la traduction automatique ont fait des progrès significatifs au cours des 20 dernières années, les résultats pour la traduction de langues morphologiquement riches sont toujours mitigés par rapport aux précédentes générations de systèmes à base de règles. Les recherches actuelles en traduction statistique de langues morphologiquement riches varient grandement en fonction de la quantité de connaissances linguistiques utilisées et de la nature de ces connaissances. Cette variation est plus importante en langue cible (par exemple, les ressources utilisées en traduction automatique statistique respectueuse de linguistique en arabe, en français et en allemand sont très différentes). La conférence portera sur les techniques état de l'art dédiées à la tâche de traduction statistique pour une langue cible qui est morphologiquement plus riche que la langue source.

La recherche d'information s'intéresse à l'accès aux documents et une majorité de travaux dans le domaine s'appuie sur les éléments textuels de ces documents écrits en langage naturel. Les requêtes soumisses par les utilisateurs de moteurs de recherche sont également textuelles, même si elles sont très pauvres d'un point de vue linguistique. Il parait donc naturel que les travaux en recherche d'information cherchent à s'alimenter par les avancées et les résultats en traitement

automatique des langues naturelles. Malgré les espoirs déçus des années 80, l'engouement pour l'utilisation du traitement du langage naturel en recherche d'information reste intact, poussé par les nouvelles perspectives offertes. Dans cette conférence, nous balayerons les aspects de la recherche d'information qui se sont le plus appuyés sur des éléments du traitement automatique des langues naturelles. Nous présenterons en particulier quelques résultats relatifs à la reformulation automatique de requêtes, à la prédiction de la difficulté des requêtes, au résumé automatique et à la contextualisation de textes courts ainsi que les perspectives actuelles offertes en particulier par les travaux en linguistique computationnelle.

Dans le but de préserver le patrimoine amazighe et éviter qu'il soit menacé de disparition, il semble opportun de doter cette langue de moyens nécessaires pour faire face aux enjeux de l'accès au domaine de l'Information et de la Communication (TIC). Dans ce contexte, et dans la perspective de construire des outils et des ressources linguistiques pour le traitement automatique de cette langue, nous avons entrepris de construire un système d'analyse morphologique pour l'amazighe standard du Maroc. Ce système profite des apports des modèles { états finis au sein de l'environnement linguistique de développement NooJ en faisant appel à des règles grammaticales à large couverture.

Nous décrivons dans cet article l'utilisation d'algorithmes d'inférence grammaticale pour la tâche de chunking, pour ensuite les comparer et les combiner avec des CRF (Conditional Random Fields), à l'efficacité éprouvée pour cette tâche. Notre corpus est extrait du French TreeBank. Nous proposons et évaluons deux manières différentes de combiner modèle symbolique et modèle statistique appris par un CRF et montrons qu'ils bénéficient dans les deux cas l'un de l'autre.

L'objectif principal du résumé multi-documents orienté par une thématique est de générer un résumé à partir de documents sources en réponse à une requête formulée par l'utilisateur. Cette

tâche est difficile car il n'existe pas de méthode efficace pour mesurer la satisfaction de l'utilisateur. Cela introduit ainsi une incertitude dans le processus de génération de résumé. Dans cet article, nous proposons une modélisation de l'incertitude en formulant notre système de résumé comme un processus de décision markovien partiellement observables (POMDP) car dans de nombreux domaines on a montré que les POMDP permettent de gérer efficacement les incertitudes. Des expériences approfondies sur les jeux de données du banc d'essai DUC ont démontré l'efficacité de notre approche.

Les travaux se focalisant sur la construction de thésaurus distributionnels ont montré que les relations sémantiques qu'ils recèlent sont principalement fiables pour les mots de forte fréquence. Dans cet article, nous proposons une méthode pour rééquilibrer de tels thésaurus en faveur des mots de fréquence faible sur la base d'un mécanisme d'amorçage : un ensemble d'exemples et de contre-exemples de mots sémantiquement similaires sont sélectionnés de façon non supervisée et utilisés pour entraîner un classifieur supervisé. Celui-ci est ensuite appliqué pour réordonner les voisins sémantiques du thésaurus utilisé pour sélectionner les exemples et contre-exemples. Nous montrons comment les relations entre les constituants de noms composés similaires peuvent être utilisées pour réaliser une telle sélection et comment conjuguer ce critère à un critère déjà expérimenté sur la symétrie des relations sémantiques. Nous évaluons l'intérêt de cette procédure sur un large ensemble de noms en anglais couvrant un vaste spectre de fréquence.

Nous proposons d'exploiter des méthodes du Traitement Automatique de Langues dédiées à la structuration de terminologie indépendamment dans deux langues (anglais et français) et de fusionner ensuite les résultats obtenus dans chaque langue. Les termes sont groupés en clusters grâce aux relations générées. L'évaluation de ces relations est effectuée au travers de la comparaison des clusters avec des données de référence et la baseline, tandis que la complémentarité des relations est analysée au travers de leur implication dans la création de

clusters de termes. Les résultats obtenus indiquent que : chaque langue contribue de manière équilibrée aux résultats, le nombre de relations hiérarchiques communes est plus grand que le nombre de relations synonymiques communes. Globalement, les résultats montrent que, dans un contexte cross-langue, chaque langue permet de détecter des régularités linguistiques et sémantiques complémentaires. L'union des résultats obtenus dans les deux langues améliore la qualité globale des clusters.

Identifier les sens possibles des mots du vocabulaire est un problème difficile demandant un travail manuel très conséquent. Ce travail a été entrepris pour l'anglais : le résultat est la base de données lexicale Word-Net, pour laquelle il n'existe encore que peu d'équivalents dans d'autres langues. Néanmoins, des traductions automatiques de Word-Net vers de nombreuses langues cibles existent, notamment pour le français. JAWS est une telle traduction automatique utilisant des dictionnaires et un modèle de langage syntaxique. Nous améliorons cette traduction, la complétons avec les verbes et adjectifs de Word-Net, et démontrons la validité de notre approche via une nouvelle évaluation manuelle. En plus de la version principale nommée WoNeF, nous produisons deux versions supplémentaires : une version à haute précision (93% de précision, jusqu'à 97% pour les noms), et une version à haute couverture contenant 109 447 paires (littéral, synset).

Les approches statistiques sont maintenant très répandues dans les différentes applications du traitement automatique de la langue et le choix d'une approche particulière dépend généralement de la tâche visée. Dans le cadre de l'interprétation sémantique multi-lingue, cet article présente une comparaison entre les méthodes utilisées pour la traduction automatique et celles utilisées pour la compréhension de la parole. Cette comparaison permet de proposer une approche unifiée afin de réaliser un décodage conjoint qui à la fois traduit une phrase et lui attribue ses étiquettes sémantiques. Ce décodage est obtenu par une approche à base de transducteurs à états finis qui permet de composer un graphe de traduction avec un graphe de compréhension. Cette

représentation peut être généralisée pour permettre des transmissions d'informations riches entre les composants d'un système d'interaction vocale homme-machine.

Cet article présente un système d'identification des relations discursives dites « implicites » (à savoir, non explicitement marquées par un connecteur) pour le français. Etant donné le faible volume de données annotées disponibles, notre système s'appuie sur des données étiquetées automatiquement en supprimant les connecteurs non ambigus pris comme annotation d'une relation, une méthode introduite par (Marcu et Echihabi, 2002). Comme l'ont montré (Sporleder et Lascarides, 2008) pour l'anglais, cette approche ne généralise pas très bien aux exemples de relations implicites tels qu'annotés par des humains. Nous arrivons au même constat pour le français et, partant du principe que le problème vient d'une différence de distribution entre les deux types de données, nous proposons une série de méthodes assez simples, inspirées par l'adaptation de domaine, qui visent à combiner efficacement données annotées et données artificielles. Nous évaluons empiriquement les différentes approches sur le corpus ANNODIS : nos meilleurs résultats sont de l'ordre de 45.6% d'exactitude, avec un gain significatif de 5.9% par rapport à un système n'utilisant que les données annotées manuellement.

Nous proposons une nouvelle méthode pour améliorer significativement la performance des modèles à paires de mentions pour la résolution de la co-référence. Étant donné un ensemble d'indicateurs, notre méthode apprend à séparer au mieux des types de paires de mentions en classes d'équivalence, chacune de celles-ci donnant lieu à un modèle de classification spécifique. La procédure algorithmique proposée trouve le meilleur espace de traits (créé à partir de combinaisons de traits élémentaires et d'indicateurs) pour discriminer les paires de mentions co-référentielles. Bien que notre approche explore un très vaste ensemble d'espaces de trait, elle reste efficace en exploitant la structure des hiérarchies construites à partir des indicateurs. Nos expériences sur les données anglaises de la CoNLL-2012 Shared Task indiquent que notre

méthode donne des gains de performance par rapport au modèle initial utilisant seulement les traits élémentaires, et ce, quelque soit la méthode de formation des chaînes ou la métrique d'évaluation choisie. Notre meilleur système obtient une moyenne de 67.2 en F1-mesure MUC, B3 et CEAF ce qui, malgré sa simplicité, le situe parmi les meilleurs systèmes testés sur ces données.

Ce travail s'inscrit dans le cadre de la construction et l'enrichissement d'ontologies à partir de textes de type encyclopédique ou scientifique. L'originalité de notre travail réside dans l'extraction de relations sémantiques exprimées au-delà de la linéarité du texte. Pour cela, nous nous appuyons sur la sémantique véhiculée par les caractères typo-dispositionels qui ont pour fonction de suppléer des formulations strictement linguistiques qui seraient plus difficilement exploitables. L'étude que nous proposons concerne les relations sémantiques portées par les structures énumératives parallèles qui, bien qu'affichant des discontinuités entre ses différents composants, présentent un tout sur le plan sémantique. Ce sont des structures textuelles qui sont propices aux relations hiérarchiques. Après avoir défini une typologie des relations portées par ce type de structure, nous proposons une approche par apprentissage visant à leur identification. Sur la base de traits incorporant informations lexico-syntaxiques et typo-dispositionnelles, les premiers résultats aboutissent à une exactitude de 61,1%.

Nous présentons l'adaptation de la base d'écrits scientifiques en ligne Scientext pour un « nouveau » public : chercheurs et autres auteurs français d'écrits scientifiques, ayant besoin de rédiger en anglais. Cette adaptation a consisté à ajouter dans la base des requêtes précodées qui permettent d'afficher les contextes dans lesquels les auteurs d'articles scientifiques en anglais expriment leur objectif de recherche et à enrichir l'interface ScienQuest de nouvelles fonctionnalités pour mémoriser et réafficher les contextes pertinents, pour faciliter la consultation par un public plus large. Les nombreuses descriptions linguistiques de la rhétorique des articles scientifiques insistent sur l'importance de la création et de l'occupation d'une « niche » de recherche. Chercheurs et

doctorants ont ici un moyen d'en visualiser des exemples sans connaître sa formulation a priori, via nos requêtes. Notre évaluation sur le corpus de test en donne une précision globale de 86,5 %.

Cet article étudie la possibilité de créer un nouveau corpus écrit en français annoté morpho-syntaxiquement à partir d'un corpus annoté existant. Nos objectifs sont de se libérer de la licence d'exploitation contraignante du corpus d'origine et d'obtenir une modernisation perpétuelle des textes. Nous montrons qu'un corpus pré-annoté automatiquement peut permettre d'entraîner un étiqueteur produisant des performances état-de-l'art, si ce corpus est suffisamment grand.

Nous présentons les premiers résultats d'un corpus arboré pour le français parlé. Il a été réalisé dans le cadre du projet ANR Etape (resp. G. Gravier) en 2011 et 2012. Contrairement à d'autres langues comme l'anglais (voir le Switchboard treebank de (Meteer, 1995)), il n'existe pas de grand corpus oral du francais annoté et validé pour les constituants et les fonctions syntaxiques. Nous souhaitons construire une ressource comparable, qui serait une extension naturelle du Corpus arboré de Paris 7 (FTB : (Abeillé et al., 2003))) basé sur des textes du journal Le Monde. Nous serons ainsi en mesure de comparer, avec des annotations comparables, l'écrit et l'oral. Les premiers résultats, qui consistent à réutiliser l'analyseur de (Petrov et al., 2006) entraîné sur l'écrit, avec une phase de correction manuelle, sont encourageants.

L'analyse syntaxique (ou parsing) en dépendances par transitions se fait souvent de façon déterministe, où chaque étape du parsing propose une seule solution comme entrée de l'étape suivante. Il en va de même pour la chaîne complète d'analyse qui transforme un texte brut en graphe de dépendances, généralement décomposé en quatre modules (segmentation en phrases, en mots, étiquetage et parsing) : chaque module ne fournit qu'une seule solution au module suivant. On sait cependant que certaines ambiguïtés ne peuvent pas être levées sans prendre en considération le niveau supérieur. Dans cet article, nous présentons l'analyseur Talismane, outil

libre et complet d'analyse syntaxique probabiliste du français, et nous étudions plus précisément l'apport d'une recherche par faisceau (beam search) à l'analyse syntaxique. Les résultats nous permettent à la fois de dégager la taille de faisceau la plus adaptée (qui permet d'atteindre un score de 88,5 % d'exactitude, légèrement supérieur aux outils comparables), ainsi que les meilleures stratégies concernant sa propagation.

L'identification d'une structure thématique dans des données textuelles quelconques est une tâche difficile. La plupart des techniques existantes reposent soit sur la maximisation d'une mesure de cohésion lexicale au sein d'un segment, soit sur la détection de ruptures lexicales. Nous proposons une nouvelle technique combinant ces deux critères de manière à obtenir le meilleur compromis entre cohésion et rupture. Nous définissons un nouveau modèle probabiliste, fondé sur l'approche proposée par Utiyama & Isahara (2001), en préservant les propriétés d'indépendance au domaine et de faible a priori de cette dernière. Des évaluations sont menées sur des textes écrits et sur des transcriptions automatiques de la parole à la télévision, transcriptions qui ne respectent pas les normes des textes écrits, ce qui accroît la difficulté. Les résultats expérimentaux obtenus démontrent la pertinence de la combinaison des critères de cohésion et de rupture.

L'étude de phénomènes d'ellipses dans les modèles de l'interface syntaxe-sémantique pose certains problèmes du fait que le matériel linguistique effacé au niveau phonologique est néanmoins présent au niveau sémantique. Tel est le cas d'une ellipse verbale ou d'une élision du sujet, par exemple, phénomènes qui interviennent lorsque deux phrases reliées par une conjonction partagent le même verbe, ou le même sujet. Nous proposons un traitement de ces phénomènes dans le formalisme des grammaires catégorielles abstraites selon un patron que nous intitulons extraction/instanciation et que nous implémentons de deux manières différentes dans les ACGs.

Nous proposons dans cet article d'intégrer la notion de chunk au sein d'une architecture globale de

traitement de la phrase. Les chunks jouent un rôle important dans les théories cognitives comme ACT-R (Anderson et al., 2004) : il s'agit d'unités de traitement globales auxquelles il est possible d'accéder directement via des buffers en mémoire à court ou long terme. Ces chunks sont construits par une fonction d'activation (processus cognitif pouvant être quantifié) s'appuyant sur l'évaluation de leur relation au contexte. Nous proposons une interprétation de cette théorie appliquée à l'analyse syntaxique. Un mécanisme de construction des chunks est proposé. Nous développons pour cela une fonction d'activation tirant parti de la représentation de l'information linguistique sous forme de contraintes. Cette fonction permet de montrer en quoi les chunks sont faciles à construire et comment leur existence facilite le traitement de la phrase. Plusieurs exemples sont proposés, illustrant cette hypothèse de facilitation.

La caractérisation du contexte des mots constitue le coeur de la plupart des méthodes d'extraction de lexiques bilingues à partir de corpus comparables. Dans cet article, nous revisitons dans un premier temps les deux principales stratégies de représentation contextuelle, à savoir celle par fenêtre ou sac de mots et celle par relations de dépendances syntaxiques. Dans un second temps, nous proposons deux nouvelles approches qui exploitent ces deux représentations de manière conjointe. Nos expériences montrent une amélioration significative des résultats sur deux corpus de langue de spécialité.

Les tâches de découverte de connaissances ont pour but de faire émerger des groupes d'entités cohérents. Ils reposent le plus souvent sur du clustering, tout l'enjeu étant de définir une notion de similarité pertinentes entre ces entités. Dans cet article, nous proposons de détourner les champs aléatoires conditionnels (CRF), qui ont montré leur intérêt pour des tâches d'étiquetage supervisées, pour calculer indirectement ces similarités sur des séquences de textes. Pour cela, nous générons des problèmes d'étiquetage factices sur les données à traiter pour faire apparaître des régularités dans les étiquetages des entités. Nous décrivons comment ce cadre peut être mis en oeuvre et

l'expérimentons sur deux tâches d'extraction d'informations. Les résultats obtenus démontrent l'intérêt de cette approche non-supervisée, qui ouvre de nombreuses pistes pour le calcul de similarités dans des espaces de représentations complexes de séquences.

Cet article aborde la problématique de l'annotation automatique d'un corpus d'apprenants d'anglais. L'objectif est de montrer qu'il est possible d'utiliser un étiqueteur PoS pour annoter un corpus d'apprenants afin d'analyser les erreurs faites par les apprenants. Cependant, pour permettre une analyse suffisamment fine, des étiquettes fonctionnelles spécifiques aux phénomènes linguistiques à étudier sont insérées parmi celles de l'étiqueteur. Celui-ci est entraîné avec ce jeu d'étiquettes étendu sur un corpus de natifs avant d'être appliqué sur le corpus d'apprenants. Dans cette expérience, on s'intéresse aux usages erronés de this et that par les apprenants. On montre comment l'ajout d'une couche fonctionnelle sous forme de nouvelles étiquettes pour ces deux formes, permet de discriminer des usages variables chez les natifs et nonnatifs et, partant, d'identifier des schémas incorrects d'utilisation. Les étiquettes fonctionnelles éclairent sur le fonctionnement discursif.

Cet article présente GLÀFF, un lexique du français à large couverture extrait du Wiktionnaire, le dictionnaire collaboratif en ligne. GLÀFF contient pour chaque entrée une description morpho-syntaxique et une transcription phonémique. Il se distingue des autres lexiques existants principalement par sa taille, sa licence libre et la possibilité de le faire évoluer de façon constante. Nous décrivons ici comment nous l'avons construit, puis caractérisé en le comparant à différentes ressources connues. Cette comparaison montre que sa taille et sa qualité font de GLÀFF un candidat sérieux comme nouvelle ressource standard pour le TAL, la linguistique et la psycholinguistique.

Cet article porte sur la mise en oeuvre etsur l'étudede techniques d'extraction de relations

sémantiques à partir d'un corpus multi-lingue aligné, en vue de construire une ressource lexicale pour l'arabe. Ces relations sontextraites par transitivité de l'équivalence traductionnelle, deux lexèmes qui possèdent les mêmes équivalents dans une langue cible étant susceptibles de partager un même sens. A partir d'équivalences extraites d'un corpus multi-lingue aligné, nous tâchons d'extraire des "cliques", ou sous-graphes maximaux complets connexes, dont toutes les unités sont en interrelation, du fait d'une probable intersection sémantique. Ces cliques présentent l'intérêt de renseigner à la fois sur la synonymie et la polysémie des unités, et d'apporter une forme de désambiguïsation sémantique. Ensuite nous tâchons de relier ces cliques avec un lexique sémantique (de type Wordnet) afin d'évaluer la possibilité de récupérer pour les unités arabes des relations sémantiques définies pour des unités en d'autres langues (français, anglais ou espagnol). Les résultats sont encourageants, et montrent qu'avec des corpus adaptés ces relations pourraient permettrede construire automatiquement un réseau utile pour certaines applications de traitement de la langue arabe.

Dans cet article, nous extrayons des adjectifs relationnels français et nous les alignons automatiquement avec les noms dont ils sont dérivés en utilisant un corpus monolingue. Les alignements adjectif-nom seront ensuite utilisés dans la traduction compositionelle des termes complexes de la forme [N AdjR] à partir d'un corpus comparable français-anglais. Un nouveau terme [N N0] (ex. cancer du poumon) sera obtenu en remplaçant l'adjectif relationnel Ad jR (ex. pulmonaire) dans [N AdjR] (ex. cancer pulmonaire) par le nom N0 (ex. poumon) avec lequel il est aligné. Si aucune traduction n'est proposée pour [N AdjR], nous considérons que ses traduction(s) sont équivalentes à celle(s) de sa paraphrase [N N0]. Nous expérimentons avec un corpus comparable dans le domaine de cancer du sein, et nous obtenons des alignements adjectif-nom qui aident à traduire des termes complexes de la forme [N AdjR] vers l'anglais avec une précision de 86 %.

Cet article présente une nouvelle méthode visant à améliorer les résultats de l'approche standard utilisée pour l'extraction de lexiques bilingues à partir de corpus comparables spécialisés. Nous tentons de résoudre le problème de la polysémie des mots dans les vecteurs de contexte par l'introduction d'un processus de désambiguïsation sémantique basé sur Word-Net. Pour traduire les vecteurs de contexte, au lieu de considérer toutes les traductions proposées par le dictionnaire bilingue, nous n'utilisons que les mots caractérisant au mieux les contextes en langue cible. Les expériences menées sur deux corpus comparables spécialisés français-anglais (financier et médical) montrent que notre méthode améliore les résultats de l'approche standard plus particulièrement lorsque plusieurs mots du contexte sont ambigus.

La construction et la validation des réseaux lexico-sémantiques est un enjeu majeur en TAL. Indépendamment des stratégies de construction utilisées, inférer automatiquement de nouvelles relations à partir de celles déjà existantes est une approche possible pour améliorer la couverture et la qualité globale de la ressource. Dans ce contexte, le moteur d'inférences a pour but de formuler de nouvelles conclusions (c'est-à-dire des relations entre les termes) à partir de prémisses (des relations préexistantes). L'approche que nous proposons est basée sur une méthode de triangulation impliquant la transitivité sémantique avec un mécanisme de blocage pour éviter de proposer des relations douteuses. Les relations inférées sont proposées aux contributeurs pour être validées. Dans le cas d'invalidation, une stratégie de réconciliation est engagée pour identifier la cause de l'inférence erronée : une exception, une erreur dans les prémisses, ou une confusion d'usage causée par la polysémie.

Beaucoup des recherches menées en extraction d'information non supervisée se concentrent sur l'extraction des relations et peu de travaux proposent des méthodes pour organiser les relations extraites. Nous présentons dans cet article une méthode de clustering en deux étapes pour regrouper des relations sémantiquement équivalentes : la première étape regroupe des relations

proches par leur expression tandis que la seconde fusionne les premiers clusters obtenus sur la base d'une mesure de similarité sémantique. Nos expériences montrent en particulier que les mesures distributionnelles permettent d'obtenir pour cette tâche de meilleurs résultats que les mesures utilisant Word-Net. Nous montrons également qu'un clustering à deux niveaux permet non seulement de limiter le nombre de similarités sémantiques à calculer mais aussi d'améliorer la qualité des résultats du clustering.

La variation du sens des mots en contexte nous a conduit à enrichir le système de types utilisés dans notre analyse syntaxico-sémantique du français basé sur les grammaires catégorielles et la sémantique de Montague (ou la lambda-DRT). L'avantage majeur d'une telle sémantique profonde est de représenter le sens par des formules logiques aisément exploitables, par exemple par un moteur d'inférence. Déterminants et quantificateurs jouent un rôle fondamental dans la construction de ces formules, et il nous a fallu leur trouver des termes sémantiques adaptés à ce nouveau cadre. Nous proposons une solution inspirée des opérateurs epsilon et tau de Hilbert, éléments génériques qui s'apparentent à des fonctions de choix. Cette modélisation unifie le traitement des différents types de déterminants et de quantificateurs et autorise le liage dynamique des pronoms. Surtout, cette description calculable des déterminants s'intègre parfaitement à l'analyseur à large échelle du français Grail, tant en théorie qu'en pratique.

Cet article aborde une question centrale de l'alignement automatique, celle du diagnostic de parallélisme des documents à aligner. Les recherches en la matière se sont jusqu'alors concentrées sur l'analyse de documents parallèles par nature : corpus de textes réglementaires, documents techniques ou phrases isolées. Les phénomènes d'inversions et de suppressions/ajouts pouvant exister entre les différentes versions d'un document sont ainsi souvent ignorées. Nous proposons donc une méthode pour diagnostiquer en contexte des zones parallèles à l'intérieur des documents. Cette méthode permet la détection d'inversions ou de suppressions entre les documents à aligner.

Elle repose sur l'affranchissement de la notion de mot et de phrase, ainsi que sur la prise en compte de la Mise en Forme Matérielle du texte (MFM). Sa mise en oeuvre est basée sur des similitudes de répartition de chaînes de caractères répétées dans les différents documents. Ces répartitions sont représentées sous forme de matrices et l'identification des zones parallèles est effectuée à l'aide de méthodes de traitement d'image.

Le développement d'outils de TAL pour les dialectes de l'arabe se heurte à l'absence de ressources pour ces derniers. Comme conséquence d'une situation de diglossie, il existe une variante de l'arabe, l'arabe moderne standard, pour laquelle de nombreuses ressources ont été développées et ont permis de construire des outils de traitement automatique de la langue. Etant donné la proximité des dialectes de l'arabe, le tunisien dans notre cas, avec l'arabe moderne standard, une voie consiste à réaliser une traduction surfacique du dialecte vers l'arabe moderne standard afin de pouvoir utiliser les outils existants pour l'arabe standard. Nous décrivons dans cet article une architecture pour une telle traduction et nous l'évaluons sur les verbes.

L'incomplétude lexicale est un problème récurrent lorsque l'on cherche à traiter le langage naturel dans sa variabilité. Effectivement, il semble aujourd'hui nécessaire de vérifier et compléter régulièrement les lexiques utilisés par les applications qui analysent d'importants volumes de textes. Ceci est plus particulièrement vrai pour les flux textuels en temps réel. Dans ce contexte, notre article présente des solutions dédiées au traitement des mots inconnus d'un lexique. Nous faisons une étude des néologismes (linguistique et sur corpus) et détaillons la mise en oeuvre de modules d'analyse dédiés à leur détection et à l'inférence d'informations (forme de citation, catégorie et classe flexionnelle) à leur sujet. Nous y montrons que nous sommes en mesure, grâce notamment à des modules d'analyse des dérivés et des composés, de proposer en temps réel des entrées pour ajout aux lexiques avec une bonne précision.

Ces dernières décennies, l'accroissement des volumes de données a rendu disponible une diversité toujours plus importante de types de contenus échangés (texte, image, audio, vidéo, SMS, tweet, données statistiques, spatiales, etc.). En conséquence, de nouvelles problématiques ont vu le jour, dont la recherche d'information au sein de données potentiellement bruitées. Dans cet article, nous nous penchons sur la reconnaissance d'entités nommées au sein de transcriptions (manuelles ou automatiques) d'émissions radiodiffusées et télévisuelles. À cet effet, nous mettons en oeuvre une approche originale par fouille de données afin d'extraire des motifs, que nous nommons règles d'annotation. Au sein d'un modèle, ces règles réalisent l'annotation automatique de transcriptions. Dans le cadre de la campagne d'évaluation Etape, nous mettons à l'épreuve le système implémenté, mXS, étudions les règles extraites et rapportons les performances du système. Il obtient de bonnes performances, en particulier lorsque les transcriptions sont bruitées.

La segmentation d'un texte en Unités Discursives Minimales (UDM) a pour but de découper le texte en segments qui ne se chevauchent pas. Ces segments sont ensuite reliés entre eux afin de construire la structure discursive d'un texte. La plupart des approches existantes utilisent une analyse syntaxique extensive. Malheureusement, certaines langues ne disposent pas d'analyseur syntaxique robuste. Dans cet article, nous étudions la faisabilité de la segmentation discursive de textes arabes en nous basant sur une approche d'apprentissage supervisée qui prédit les UDM et les UDM imbriqués. La performance de notre segmentation a été évaluée sur deux genres de corpus : des textes de livres de l'enseignement secondaire et des textes du corpus Arabic Treebank. Nous montrons que la combinaison de traits typographiques, morphologiques et lexicaux permet une bonne reconnaissance des bornes de segments. De plus, nous montrons que l'ajout de traits syntaxiques n'améliore pas les performances de notre segmentation.

Une faiblesse des systèmes de traduction statistiques est le caractère ad hoc du processus d'apprentissage, qui repose sur un empilement d'heuristiques et conduit à apprendre des

paramètres dont la valeur est sous-optimale. Dans ce travail, nous reformulons la traduction automatique sous la forme familière de l'apprentissage d'un modèle probabiliste structuré utilisant une paramétrisation log-linéaire. Cette entreprise est rendue possible par le développement d'une implantation efficace qui permet en particulier de prendre en compte la présence de variables latentes dans le modèle. Notre approche est comparée, avec succès, avec une approche de l'état de l'art sur la tâche de traduction de données du BTEC pour le couple Français-Anglais.

Explorer et maintenir une documentation technique est une tâche difficile pour laquelle on pourrait bénéficier d'un outillage efficace, à condition que les documents soient annotés sémantiquement. Les annotations doivent être riches, cohérentes, suffisamment spécialisées et s'appuyer sur un modèle sémantique explicite ? habituellement une ontologie ? qui modélise la sémantique du domaine cible. Il s'avère que les approches d'annotation traditionnelles donnent pour cette tâche des résultats limités. Nous proposons donc une nouvelle approche, l'annotation sémantique statistique basée sur les syntagmes, qui prédit les annotations sémantiques à partir d'un ensemble d'apprentissage réduit. Cette modélisation facilite l'annotation sémantique spécialisée au regard de modèles sémantiques de domaine arbitrairement riches. Nous l'évaluons à l'aide de plusieurs métriques et sur deux textes décrivant des réglementations métier. Notre approche obtient de bons résultats. En particulier, la F-mesure est de l'ordre de 91, 9% et 97, 6% pour la prédiction de l'étiquette et de la position avec différents paramètres. Cela suggère que les annotateurs humains peuvent être fortement aidés pour l'annotation sémantique dans des domaines spécifiques.

Dans cet article, nous présentons une méthode de segmentation de pages web en blocs de texte pour la sélection de documents pertinents en questions-réponses. La segmentation des documents se fait préalablement à leur indexation en plus du découpage des segments obtenus en passages au moment de l'extraction des réponses. L'extraction du contenu textuel des pages est faite à l'aide d'un extracteur maison. Nous avons testé deux méthodes de segmentation. L'une segmente les

textes extraits des pages web uniformément en blocs de taille fixe, l'autre les segmente par TextTiling (Hearst, 1997) en blocs thématiques de taille variable. Les expériences menées sur un corpus de 500K pages web et un jeu de 309 questions factuelles en français, issus du projet Quaero (Quintard et al., 2010), montrent que la méthode employée tend à améliorer la précision globale (top-10) du système RITEL?QR (Rosset et al., 2008) dans sa tâche.

La simplification lexicale consiste à remplacer des mots ou des phrases par leur équivalent plus simple. Dans cet article, nous présentons trois modèles de simplification lexicale, fondés sur différents critères qui font qu'un mot est plus simple à lire et à comprendre qu'un autre. Nous avons testé différentes tailles de contextes autour du mot étudié : absence de contexte avec un modèle fondé sur des fréquences de termes dans un corpus d'anglais simplifié ; quelques mots de contexte au moyen de probabilités à base de n-grammes issus de données du web ; et le contexte étendu avec un modèle fondé sur les fréquences de co-occurrences.

Cet article présente une méthode générative de prédiction de la structure sémantique en cadres d'une phrase à partir de sa structure syntaxique et décrit les grammaires utilisées ainsi que leurs performances. Ce système permet de prédire, pour un mot dans le contexte syntaxique d'une phrase, le cadre le plus probable. Le système génératif permet d'attribuer à ce mot un cadre et à l'ensemble de chemins des rôles sémantiques. Bien que les résultats ne soient pas encore satisfaisants, cet analyseur permet de regrouper les tâches d'analyse sémantique (sélection du cadre, sélection des actants, attribution des rôles), contrairement aux travaux précédemment publiés. De plus, il offre une nouvelle approche de l'analyse sémantique en cadres, dans la mesure où elle repose plus sur la structure syntaxique que sur les mots de la phrase.

Dans le cadre du projet ASFALDA, qui comporte une phase d'annotation sémantique d'un FrameNet français, nous cherchons à fournir un traitement linguistiquement motivé des

constructions à attribut de l'objet, un exemple typique de divergence syntaxe-sémantique. Pour ce faire, nous commençons par dresser un panorama des propriétés syntaxiques et sémantiques des constructions à attribut de l'objet. Nous étudions ensuite le traitement FrameNet des verbes anglais typiques de cette construction, avant de nous positionner pour un traitement homogénéisé dans le cas du FrameNet français.

Devant des collections massives et hétérogènes de données, les systèmes de RI doivent désormais pouvoir appréhender des comportements d'utilisateurs aussi variés qu'imprévisibles. L'objectif de notre approche est d'évaluer la façon dont un utilisateur verbalise un besoin informationnel à travers un énoncé de type « expression libre » ; appelé langage naturel (LN). Pour cela, nous nous situons dans un contexte applicatif, à savoir des demandes de remboursement des utilisateurs d'un moteur de recherche dédié à des études économiques en français. Nous avons recueilli via ce moteur, les demandes en LN sur 5 années consécutives totalisant un corpus de 1398 demandes. Nous avons alors comparé l'expression en tant que tel du besoin informationnel en fonction de la tâche de recherche d'informations (RI) de l'utilisateur.

Dans le cadre de notre projet de recherche, qui a pour but l'implémentation d'un outil de simplification des emplois spécialisés de verbes dans des corpus médicaux à partir de l'analyse syntaxico-sémantique de ces verbes en contexte, nous proposons une analyse de quelques approches et travaux qui ont pour objet principal la description du verbe dans les trois domaines de recherche à l'interface desquels se situe notre projet : linguistique, TAL et terminologie. Nous décrivons plus particulièrement les travaux qui peuvent avoir une incidence sur notre étude. Cet état de l'art nous permet de mieux connaître le cadre théorique dans lequel s'intègre notre projet de recherche et d'avoir les repères et références susceptibles de contribuer à sa réalisation.

L'analyse syntaxique et sémantique de langages non-canoniques est principalement limitée par le

manque de corpus annotés. Il est donc primordial de mettre au point des systèmes robustes capables d'allier références canoniques et non-canoniques. Les méthodes exploitant la théorie des réseaux de neurones profonds ont prouvé leur efficacité dans des domaines tels que l'imagerie ou les traitements acoustiques. Nous proposons une architecture de réseau de neurones appliquée au traitement automatique des langages naturels, et plus particulièrement à l'étiquetage morpho-syntaxique. De plus, plutôt que d'extraire des représentations empiriques d'une phrase pour les injecter dans un algorithme de classification, nous nous inspirons de récents travaux portant sur l'extraction automatique de représentations vectorielles des mots à partir de corpus non-annotés. Nous souhaitons ainsi tirer profit des propriétés de linéarité et de compositionnalité de tels plongements afin d'améliorer les performances de notre système.

Cet article présente une méthode ayant pour objectif de minimiser l'apport extérieur nécessaire à la tâche d'extraction terminologique (ET) et de rendre cette tâche moins dépendante de la langue. Pour cela, la méthode prévoit des ressources morphologiques et morpho-syntaxiques simplifiées construites directement à partir d'un corpus lemmatisé. Ces ressources endogènes servent à la création d'un système de filtres qui affinent les calculs statistiques et à la génération de patrons pour l'identification de candidats termes polylexicaux. La méthode a été testée sur deux corpus comparables en chimie et en télécommunication, en français et en anglais. La précision observée sur les 100 premiers candidats termes monolexicaux fluctue entre 71% et 87% pour le français et entre 44 % et 69 % en anglais ; celle des candidats termes polylexicaux s'élève à 69-78 % en français et 69-85 % en anglais en fonction du domaine.

Dans cet article, nous abordons la problématique du fonctionnement de la temporalité en langue des signes française (LSF). Nous allons étudier plus particulièrement quelques structures portant sur la durée. Nous présenterons dans un premier temps les descriptions existantes du système aspecto-temporel de la LSF et les difficultés que nous trouvons pour modéliser ces travaux. Le but

de cet article est de proposer une grammaire formelle qui prenne en compte le fonctionnement de la LSF et qui puisse faire l'objet d'un traitement de modélisation. Notre démarche consiste à étudier un corpus LSF pour établir des liens de fonction à forme afin d'obtenir des règles de grammaire qu'on peut générer dans un projet de synthèse à l'aide d'un signeur avatar.

Cet article présente le problème de l'association entre énoncés en langage naturel exprimant des instructions opérationnelles et leurs expressions équivalentes et langage formel. Nous l'appliquons au cas du français et du langage R. Développer un assistant opérationnel apprenant, qui constitue notre objectif à long terme, requiert des moyens pour l'entraîner et l'évaluer, c'est-à-dire un système initial capable d'interagir avec l'utilisateur. Après avoir introduit la ligne directrice de ce travail, nous proposons un modèle pour représenter le problème et discutons de l'adéquation des méthodes par mise en correspondance, ou mapping, à notre tâche. Pour finir, nous montrons que, malgré des scores modestes, une approche simple semble suffisante pour amorcer un tel système interactif apprenant.

Cet article propose une méthode de regroupement de structures de dérivations lexicales par raisonnement analogique. Nous présentons les caractéristiques générales d'un graphe lexical issu du Réseau Lexical du Français, dont nous exploitons par la suite les composantes faiblement connexes. Ces composantes sont regroupées en trois étapes : par isomorphisme, par similarité de relations, puis par similarité d'attributs. Les résultats du dernier regroupement sont analysés en détail.

Dans ce papier nous traitons des résumés automatiques de conversations parlées spontanées. Pour cela nous utilisons des conversations provenant de cas réels d'appels téléphoniques de centre d'appels issues du corpus DECODA. Nous testons des méthodes extractives classiques utilisées en résumé de texte (MMR) ainsi que des méthodes basées sur des heuristiques du dialogue dans le

cadre des centres d'appels. Il s'agit de la sélection du tour de parole le plus long dans le premier quart de la conversation, dans l'ensemble de la conversation et dans le dernier quart de la conversation. L'ensemble est évalué avec la métrique ROUGE. Les résultats obtenus soulignent les limites de ces approches « classiques » et confirment la nécessité d'envisager des méthodes abstractives intégrant des informations de structures sur les conversations. En effet, ces premiers résultats montrent que les méthodes heuristiques basées sur la structure produisent des résultats comparables, voir meilleurs que des méthodes telles que MMR.

Dans les études s'intéressant à la traduction assistée par ordinateur (TAO), l'un des objectifs consiste à repérer les passages qui focalisent l'attention des traducteurs humains. Ces passages sont considérés comme étant des contextes « riches en connaissances », car ils aident à la traduction. Certains contextes textuels ne donnent qu'une simple attestation d'un terme recherché, même s'ils le renferment. Ils ne fournissent pas d'informations qui permettent de saisir sa signification. D'autres, en revanche, contiennent des fragments de définitions, mentionnent une variante terminologique ou utilisent d'autres notions facilitant la compréhension du terme. Ce travail s'intéresse aux « contextes définitoires » qui sont utiles à l'acquisition de connaissances à partir de textes, en particulier dans la perspective de traduction terminologique assistée par ordinateur à partir de corpus comparables. En effet, l'association d'un exemple à un terme permet d'en appréhender le sens exact. Nous proposons, tout d'abord, trois hypothèses définissant la notion d'exemple définitoire. Ensuite nous évaluons sa validité grâce une méthode s'appuyant sur les Contextes Riches en Connaissances (CRC) ainsi que les relations hiérarchiques reliant les termes entre eux.

Dans cet article, nous présentons une démarche pour l'induction d'une grammaire de propriétés (GP) arabe en utilisant le treebank ATB. Cette démarche se base sur deux principales étapes : (1) l'induction d'une grammaire hors contexte et (2) l'induction d'une GP par la génération automatique

des relations qui peuvent exister entre les unités grammaticales décrites dans la CFG. Le produit obtenu constitue une ressource ouvrant de nouvelles perspectives pour la description et le traitement de la langue arabe.

Les techniques actuelles de traduction automatique (TA) permettent de produire des traductions dont la qualité ne cesse de croitre. Dans des domaines spécifiques, la post-édition (PE) de traductions automatiques permet, par ailleurs, d'obtenir des traductions de qualité relativement rapidement. Mais un tel pipeline (TA+PE) est il envisageable pour traduire une oeuvre littéraire ? Cet article propose une ébauche de réponse à cette question. Un essai de l'auteur américain Richard Powers, encore non disponible en français, est traduit automatiquement puis post-édité et révisé par des traducteurs non-professionnels. La plateforme de post-édition du LIG utilisée permet de lire et éditer l'oeuvre traduite en français continuellement, suggérant (pour le futur) une communauté de lecteurs-réviseurs qui améliorent en continu les traductions de leur auteur favori. En plus de la présentation des résultats d'évaluation expérimentale du pipeline TA+PE (système de TA utilisé, scores automatiques), nous discutons également la qualité de la traduction produite du point de vue d'un panel de lecteurs (ayant lu la traduction en français, puis répondu à une enquête). Enfin, quelques remarques du traducteur français de R. Powers, sollicité à cette occasion, sont présentées à la fin de cet article.

Dans cet article, nous montrons comment l'utilisation conjointe d'une technique d'alignement de phrases parallèles à la demande et d'estimation de modèles de traduction à la volée permet une réduction en temps très notable (jusqu'à 93% dans nos expériences) par rapport à un système à l'état de l'art, tout en offrant un compromis en termes de qualité très intéressant dans certaines configurations. En particulier, l'exploitation immédiate de documents traduits permet de compenser très rapidement l'absence d'un corpus de développement.

Nous présentons un projet collaboratif en cours mené par l'université de Grenoble et l'université de Xiamen, et visant à créer des instances d'un nouveau type de système de traduction automatique statistique utilisant des ressources lexico-sémantiques et discursives. Le but concret est de développer des systèmes de TAS chinois-français pour des sites boursiers et économiques. Comme très peu de corpus et de dictionnaires bilingues chinois-français sont disponibles sur Internet, l'anglais est utilisé comme "pivot" pour construire les équivalents chinois-français par transitivité. Outre la description générale de ce projet, nous décrivons les progrès sur deux axes de recherche liés à ce projet. Pour cela, nous utilisons une méthode, proposée par XMU, d'induction de probabilité fondée sur la similarité thématique, qui produit des tables de traduction C-F à partir de tables de traduction C-E et E-F. Pour disposer de bons corpus parallèles C-F, nous utilisons un système Web de post-édition collaborative qui peut déclencher l'amélioration incrémentale du système de TA en utilisant des métriques d'évaluation de TA et en extrayant la "meilleure partie" de la mémoire de traductions courante.

Pour une communauté, la terminologie est essentielle car elle permet de décrire, échanger et récupérer les données. Dans de nombreux domaines, l'explosion du volume des données textuelles nécessite de recourir à une automatisation du processus d'extraction de la terminologie, voire son enrichissement. L'extraction automatique de termes peut s'appuyer sur des approches de traitement du langage naturel. Des méthodes prenant en compte les aspects linguistiques et statistiques proposées dans la littérature, résolvent quelques problèmes liés à l'extraction de termes tels que la faible fréquence, la complexité d'extraction de termes de plusieurs mots, ou l'effort humain pour valider les termes candidats. Dans ce contexte, nous proposons deux nouvelles mesures pour l'extraction et le "ranking" des termes formés de plusieurs mots à partir des corpus spécifiques d'un domaine. En outre, nous montrons comment l'utilisation du Web pour évaluer l'importance d'un terme candidat permet d'améliorer les résultats en terme de précision. Ces expérimentations sont réalisées sur le corpus biomédical GENIA en utilisant des mesures de la littérature telles que

C-value.

Basé sur les calculs d'entropie conditionnelle de (Bonami & Boyé, à paraître), nous proposons un analyseur automatique de la flexion dans le cadre de la morphologie thématique qui produit le graphe de régularités du paradigme. Le traitement se base sur un lexique de 6440 verbes extraits du BDLex (de Calmès & Pérennou, 1998) associés à leurs fréquences dans Lexique3 (New et al., 2001). L'algorithme se compose de trois éléments : calcul de l'entropie conditionnelle entre paires de formes fléchies, distillation des paradigmes, construction du graphe de régularités. Pour l'entropie, nous utilisons deux modes de calcul différents, l'un se base sur la distribution de l'effectif des verbes entre leurs différentes options, l'autre sur la distribution des lexèmes verbaux en fonction de leurs fréquences pour contrebalancer l'influence des verbes ultra-fréquents sur les calculs.

Le traitement informatique de constructions à verbe support (prendre une photo, faire une présentation) est une tâche difficile en TAL. Cela est également vrai en espagnol, où ces constructions sont fréquentes dans les textes, mais ne font pas souvent partie des lexiques exploitables par une machine. Notre objectif est d'extraire des constructions à verbe support à partir d'un très grand corpus de l'espagnol. Nous peaufinons un ensemble de motifs morpho-syntaxiques fondés sur un grand nombre de verbe support possibles. Ensuite, nous filtrons cette liste en utilisant des seuils et des mesures d'association. Bien que tout à fait classique, cette méthode permet l'extraction de nombreuses expressions de bonne qualité. À l'avenir, nous souhaitons étudier les représentations sémantiques de ces constructions dans des lexiques multi-lingues.

La sous-catégorisation d'arguments introduits par la préposition pour a été sous-étudiée par le passé, comme en témoigne l'incomplétude des ressources lexico-syntaxiques existantes sur ce point. Dans cet article, nous présentons rapidement les différents types de sous-catégorisation en pour, qui contrastent avec les emplois de pour comme connecteur de discours. Nous décrivons

l'intégration des arguments en pour au lexique syntaxique Lefff , enrichissant ainsi les informations de sous-catégorisation de nombreuses entrées verbales, nominales, adjectivales et adverbiales.

Les outils d'étiquetage automatique sont plus ou moins robustes en ce qui concerne l'étiquetage de mots inconnus, non rencontrés dans le corpus d'apprentissage. Il est important de connaître de manière précise la performance de ces outils lorsqu'on cible plus particulièrement l'étiquetage de néologismes formels. En effet, la catégorie grammaticale constitue un critère important à la fois pour leur identification et leur documentation. Nous présentons une évaluation et une comparaison de 7 étiqueteurs morpho-syntaxiques du français, à partir d'un corpus issu du Wiktionnaire. Les résultats montrent que l'utilisation de traits de forme ou morphologiques est favorable à l'étiquetage correct des mots nouveaux.

La correction de données textuelles obtenues par reconnaissance optique de caractères (OCR) pour atteindre une qualité éditoriale reste aujourd'hui une tâche coûteuse, car elle implique toujours une intervention humaine. La détection et la correction automatiques d'erreurs à l'aide de modèles statistiques ne permettent de traiter de façon utile que les erreurs relevant de la langue générale. C'est pourtant dans certaines entités nommées que résident les erreurs les plus nombreuses, surtout dans des données telles que des corpus de brevets ou des textes juridiques. Dans cet article, nous proposons une architecture d'identification et de correction par règles d'un large éventail d'entités nommées (non compris les noms propres). Nous montrons que notre architecture permet d'atteindre un bon rappel et une excellente précision en correction, ce qui permet de traiter des fautes difficiles à traiter par les approches statistiques usuelles.

Nous présentons dans ce travail un nouveau système de voyellation automatique des textes arabes en utilisant trois étapes. Durant la première phase, nous avons intégré une base de données lexicale contenant les mots les plus fréquents de la langue arabe avec l'analyseur morphologique

AlKhalil Morpho Sys pour fournir les voyellations possibles pour chaque mot. Le second module dont l'objectif est d'éliminer l'ambiguïté repose sur une approche statistique dont l'apprentissage a été effectué sur un corpus constitué de textes de livres arabes et utilisant les modèles de Markov cachés (HMM) où les mots non voyellés représentent les états observés et les mots voyellés sont ses états cachés. Le système utilise les techniques de lissage pour contourner le problème des transitions des mots absentes et l'algorithme de Viterbi pour sélectionner la solution optimale. La troisième étape utilise un modèle HMM basé sur les caractères pour traiter le cas des mots non analysés.

Dans cet article, nous proposons une évaluation dans un cadre utilisateur de Citron, un système de question-réponse en français capable d'extraire des réponses à des questions à réponses multiples (questions possédant plusieurs réponses correctes différentes) en domaine ouvert à partir de documents provenant du Web. Nous présentons ici le protocole expérimental et les résultats pour nos deux expériences utilisateurs qui visent à (1) comparer les performances de Citron par rapport à celles d'un être humain pour la tâche d'extraction de réponses multiples et (2) connaître la satisfaction d'un utilisateur devant différents formats de présentation de réponses.

Dans ces travaux, nous présentons une approche afin d'étiqueter une large collection de chansons francophones. Nous avons développé une interface utilisant les paroles comme point d'entrée afin d'explorer cette collection de musique avec des filtres en fonction de chaque période musicale. Dans un premier temps, nous avons collecté paroles et méta-données de différentes sources sur leWeb. Nous présentons dans cet article une méthode originale permettant d'attribuer de manière automatique la décennie de sortie des chansons de notre collection. Basée sur un système évalué au cours d'une des campagnes DEFT, l'approche combine fouille de textes et apprentissage supervisé et aborde la problématique comme une tâche de classification multi classes. Nous avons par la suite enrichi le modèle d'un certain nombre de traits supplémentaires tels que les tags

sociaux afin d'observer leur influence sur les résultats.

Cet article présente une plateforme dédiée à l'évaluation de la difficulté des textes administratifs, dans un but d'aide à la rédaction. La plateforme propose d'une part une formule de lisibilité spécialisée pour les textes administratifs, dont la conception repose sur une nouvelle méthode d'annotation. Le modèle classe correctement 58% des textes sur une échelle à 5 niveaux et ne commet d'erreurs graves que dans 9% des cas. La plateforme propose d'autre part un diagnostic plus précis des difficultés spécifiques d'un texte, sous la forme d'indicateurs numériques, mais aussi d'une localisation de ces difficultés directement dans le document.

Les « mots dièses» ou « hash tags » sont le moyen naturel de lier entre eux différents tweets. Certains « hash tags » sont en fait de petites phrases dont la décomposition peut se révéler particulièrement utile lors d'une analyse d'opinion des tweets. Nous allons montrer dans cet article comment l'on peut automatiser cette décomposition et cette analyse de façon à améliorer la détection de la polarité des tweets.

Cet article aborde la question des expressions d'attitudes (affects, jugements, appréciations) chez l'utilisateur dans le cadre d'échanges avec un agent virtuel. Il propose une méthode pour l'analyse des réponses à des questions fermées destinée à interroger les attitudes de l'utilisateur. Cette méthode s'appuie sur une formalisation des questions de l'agent ? sous la forme d'une fiche linguistique ? et sur une analyse de la réponse de l'utilisateur, pour créer un modèle utilisateur. La fiche linguistique de l'agent est structurée par un ensemble d'attributs relatifs, d'une part, à l'attitude à laquelle réfère la question, d'autre part, à sa forme morpho-syntaxique. L'analyse de la réponse, quant à elle, repose sur un ensemble de règles sémantiques et syntaxiques définies par une grammaire formelle. A partir des résultats fournis par cette analyse et des informations contenues dans la fiche linguistique de l'agent, des calculs sémantiques sont effectués pour définir la valeur de

la réponse et construire le modèle utilisateur.

Cet article présente une bibliothèque python appelée KNG permettant d'écrire facilement des automates et transducteurs finis. Grâce à une gestion soigneuse des codages et des entrées-sorties, cette bibliothèque permet de réaliser une cascade de transducteurs au moyen de tubes unix reliant des scripts python.

L'objectif est de comparer deux outils d'analyse de corpus de textes bruts pour l'aide à la recherche en linguistique japonaise. Les deux outils représentent chacun une approche spécifique. Le premier, Sagace, recherche un patron sans prise en compte de son environnement. Le second, un dispositif à base de Mecab, recherche les patrons après analyse morphologique complète des phrases. Nous comparons les performances en temps et en précision. Il ressort de cette analyse que les performances de Sagace sont globalement un peu inférieures à celles des dispositifs à base de Mecab, mais qu'elles restent tout à fait honorables voire meilleures pour certaines tâches.

Les concordanciers jouent depuis longtemps un rôle important dans l'analyse des corpus linguistiques, tout comme dans les domaines de la philologie, de la littérature, de la traduction et de l'enseignement des langues. Toutefois, il existe peu de concordanciers qui soient capables d'associer des annotations à plusieurs niveaux et synchronisées avec le signal sonore. L'essor des grands corpus de français parlé introduit une augmentation des exigences au niveau de la performance. Dans ce travail à caractère préliminaire, nous avons développé un prototype de concordancier multi-niveaux et multimédia, que nous avons testé sur le corpus de français parlé du projet Phonologie du Français Contemporain (PFC, 1,5 million de tokens de transcription alignée au niveau de l'énoncé). L'outil permet non seulement d'enrichir les résultats des concordances grâce aux données relevant de plusieurs couches d'annotation du corpus (annotation morpho-syntaxique, lemme, codage de la liaison, codage du schwa etc.), mais aussi d'élargir les modalités d'accès au

corpus.

On présente une étude d'apprentissage visant à montrer que les contextes locaux dans un corpus de parole adressée aux enfants peuvent être exploités, avec des méthodes statistiques simples, pour prédire la catégorie (nominale vs. verbale) d'un mot inconnu. Le modèle présenté, basé sur la mémorisation de n-grammes et sur une « graine sémantique » (un petit nombre de noms et verbes supposés connus et catégorisés) montre une excellente précision à toutes les tailles de graine sémantique, et un rappel plus faible, qui croît avec la taille de la graine sémantique. Les contextes les plus utilisés sont ceux qui contiennent des mots fonctionnels. Cette étude de faisabilité démontre que les très jeunes enfants pourraient exploiter les contextes de mots inconnus pour prédire leur catégorie syntaxique.

Nous présentons une étude comparative sur l'impact de la nature et de la taille des corpus d'apprentissage sur les performances dans la détection automatique des entités nommées. Cette évaluation se présente sous la forme de multiples modulations de trois corpus français. Deux des corpus sont issus du catalogue des ressources linguistiques d'ELRA et le troisième est composé de documents extraits de la plateforme OpenEdition.org.

La reconnaissance des Entités Nommées (REN) en langue amazighe est un pré-traitement potentiellement utile pour de nombreuses applications du traitement de la langue amazighe. Cette tâche représente toutefois un sévère challenge, compte tenu des particularités de cette langue. Dans cet article, nous présentons le premier système d'extraction d'entités nommées amazighes (RENAM) fondé sur une approche symbolique qui utilise le principe de transducteur à états finis disponible sous la plateforme GATE.

Dans cet article nous employons le « topic modeling » pour explorer des chemins vers la détection

automatique de l'apparition de nouveaux sens pour des mots connus. Nous appliquons les méthodes introduites dans (Lau et al., 2012, 2014) à un cas de néologie sémantique récent, l'apparition du nouveau sens de geste pour le mot « quenelle ». Nos expériences mettent en évidence le potentiel de cette approche pour l'apprentissage des sens du mot, l'alignement des topics à des sens de dictionnaire et enfin la détection de nouveaux sens.

Cet article propose une approche pour la formalisation de grammaires pour les langues des signes, rendant compte de leurs particularités linguistiques. Comparable aux grammaires génératives en termes de récursivité productive, le système présente des propriétés nouvelles comme la multi-linéarité permettant la spécification simultanée des articulateurs. Basé sur l'analyse des liens entre formes produites/observées et fonctions linguistiques au sens large, on observe un décloisonnement des niveaux traditionnels de construction de la langue, inhérent à la méthodologie employée. Nous présentons un ensemble de règles trouvées suivant la démarche présentée et concluons avec une perspective intéressante en traduction automatique vers la langue des signes.

Les langues des signes sont les langues naturelles utilisées dans les communautés sourdes. Elles ont suscitées, ces dernières années, de l'intérêt dans le domaine du traitement automatique des langues naturelles. Néanmoins, il y a un manque général de données de terrain homogènes. Notre réflexion porte sur les moyens de soutenir la maturation des modèles linguistiques avant d'entamer un large effort d'annotation. Cet article présente des pré-requis pour la réalisation d'outils s'inscrivant dans cette démarche. L'exposé est illustré avec deux outils développés pour les langues des signes : le premier utilise une logique adaptée pour la représentation de modèles phonologiques et le second utilise des grammaires formelles pour la représentation de modèles syntaxiques.

La tâche du résumé multi-lingue vise à concevoir des systèmes de résumé très peu dépendants de

la langue. L'approche par extraction est au coeur de ces systèmes, elle permet à l'aide de méthodes statistiques de sélectionner les phrases les plus pertinentes dans la limite de la taille du résumé. Dans cet article, nous proposons une approche de résumé multi-lingue, elle extrait les phrases dont les termes sont des plus discriminants. De plus, nous étudions l'impact des différents traitements linguistiques de base : le découpage en phrases, l'analyse lexicale, le filtrage des mots vides et la racinisation sur la couverture ainsi que la notation des phrases. Nous évaluons les performances de notre approche dans un contexte multi-lingue : l'anglais, l'arabe et le français en utilisant le jeu de données TAC MultiLing 2011.

Nous présentons dans cet article l'adaptation de l'outil de résumé automatique REZIME à la langue française. REZIME est un outil de résumé automatique mono-document destiné au domaine médical et s'appuyant sur des critères statistiques, syntaxiques et lexicaux pour extraire les phrases les plus pertinentes. Nous décrivons dans cet article le système REZIME tel qu'il a été conçu et les différentes étapes de son adaptation à la langue française. Les performances de l'outil adapté au français sont mesurées et comparées à celle de la version anglaise. Les résultats montrent que l'adaptation au français ne dégrade pas les performances de REZIME, qui donne des résultats équivalents dans les deux langues.

Cet article aborde le problème de l'extraction de données orales multi-annotées : nous proposons une solution intermédiaire, entre d'une part les systèmes de requêtages très évolués mais qui nécessitent des données structurées, d'autre part les données (multi-)annotées des utilisateurs qui sont hétérogènes. Notre proposition s'appuie sur 2 fonctions principales : une fonction booléenne pour filtrer sur le contenu, et une fonction de relation qui implémente l'algèbre de Allen. Le principal avantage de cette approche réside dans sa généricité : le fonctionnement sera identique que les annotations proviennent de Praat, Transcriber, Elan ou tout autre logiciel d'annotation. De plus, deux niveaux d'utilisation ont été développés : une interface graphique qui ne nécessite aucune

compétence ou connaissance spécifique de la part de l'utilisateur, et un interrogation par scripts en langage Python. L'approche a été implémentée dans le logiciel SPPAS, distribué sous licence GPL.

Cet article propose une analyse critique de la norme TimeML à la lumière de l'expérience d'annotation temporelle d'un corpus de français parlé. Il montre que certaines adaptations de la norme seraient conseillées pour répondre aux besoins du TAL et des sciences du langage. Sont étudiées ici les questions de séparation des niveaux d'annotation, de délimitation des éventualités dans le texte et de l'ajout d'une relation temporelle de type associative.

De nombreuses informations cliniques sont contenues dans le texte des dossiers électroniques de patients et ne sont pas directement accessibles à des fins de traitement automatique. Pour pallier cela, nous préparons un large corpus annoté de documents cliniques. Une première étape de ce travail consiste à séparer le contenu médical des documents et les informations administratives contenues dans les en-têtes et pieds de page. Nous présentons un système d'identification automatique de zones dans les documents cliniques qui offre une F-mesure de 0,97, équivalente à l'accord inter-annoteur de 0,98. Notre étude montre que le contenu médical ne représente que 60% du contenu total de notre corpus, ce qui justifie la nécessité d'une segmentation en zones. Le travail d'annotation en cours porte sur les sections médicales identifiées.

À partir du schéma d'annotation en dépendances syntaxiques de surface du corpus Sequoia, nous proposons un schéma en dépendances syntaxiques profondes qui en est une abstraction exprimant les relations grammaticales entre mots sémantiquement pleins. Quand ces relations grammaticales sont partie prenante de diathèses verbales, ces diathèses sont vues comme le résultat de redistributions à partir d'une diathèse canonique et c'est cette dernière qui est retenue dans notre schéma d'annotation syntaxique profonde.

Cet article présente un essai d'application de l'analyse argumentative (text zoning) à l'ACL Anthology. Il s'agit ainsi de mieux caractériser le contenu des articles du domaine de la linguistique informatique afin de pouvoir en faire une analyse fine par la suite. Nous montrons que des technique récentes d'analyse argumentative fondées sur l'apprentissage faiblement supervisé permettent d'obtenir de bons résultats.

Nous présentons un ensemble d'exemples lexicographiques intégré dans le Réseau Lexical du Français et explorons son intérêt potentiel en tant que corpus annoté pour la recherche en désambiguisation sémantique automatique.

En TAL et plus particulièrement en analyse sémantique, les informations sur la couleur peuvent être importantes pour traiter correctement des informations textuelles (sens des mots, désambiguïsation et indexation). Plus généralement, connaître la ou les couleurs habituellement associée(s) à un terme est une information cruciale. Dans cet article, nous montrons comment le crowdsourcing, à travers un jeu, peut être une bonne stratégie pour collecter ces données lexico-sémantiques.

En traitement automatique des langues, les ressources lexico-sémantiques ont été incluses dans un grand nombre d'applications. La création manuelle de telles ressources est consommatrice de temps humain et leur couverture limitée ne permet pas toujours de couvrir les besoins des applications. Ce problème est encore plus important pour les langues moins dotées que le français ou l'anglais. L'induction de sens présente dans ce cadre une piste intéressante. À partir d'un corpus de texte, il s'agit d'inférer les sens possibles pour chacun des mots qui le composent. Nous étudions dans cet article une approche basée sur une représentation vectorielle pour chaque occurrence d'un mot correspondant à ses voisins. À partir de cette représentation, construite sur un corpus en bengali, nous comparons plusieurs approches de classification non-supervisées (k-moyennes, regroupement hiérarchique et espérance-maximisation) des occurrences d'un mot pour déterminer

les différents sens qu'il peut prendre. Nous comparons nos résultats au Bangla Word-Net ainsi qu'à une référence établie pour l'occasion. Nous montrons que cette méthode permet de trouver des sens qui ne se trouvent pas dans le Bangla Word-Net.

L'article présente deux ressources pour le TAL, distribuées sous licence GPL : un dictionnaire de mots composés français et une grammaire NooJ spécifiant un sous-ensemble des schémas de composés.

Nous proposons une démonstration de la Plateforme de l'Equipex ORTOLANG (Open Resources and Tools for LANGuage : www.ortolang.fr) en cours de mise en place dans le cadre du programme d'investissements d'avenir (PIA) lancé par le gouvernement français. S'appuyant entre autres sur l'existant des centres de ressources CNRTL (Centre National de Ressources Textuelles et Lexicales : www.cnrtl.fr) et SLDR (Speech and Language Data Repository : http://sldr.org/), cette infrastructure a pour objectif d'assurer la gestion, la mutualisation, la diffusion et la pérennisation de ressources linguistiques de type corpus, dictionnaires, lexiques et outils de traitement de la langue, avec une focalisation particulière sur le français et les langues de France.

Dans ce papier, nous nous intéressons à l'utilisation d'une ressource linguistique propriétaire riche pour une tâche de classification. L'objectif est ici de mesurer l'impact de l'ajout de ces ressources sur cette tâche en termes de performances. Nous montrons que l'utilisation de cette ressource en temps que traits supplémentaires de classification apporte un réel avantage pour un ajout très modéré en termes de nombre de traits.

Cette démonstration de CFAsT s'intéresse à "comment concevoir un système de dialogue avec un effort minimal". Cet assistant virtuel repose sur un nouveau modèle pour la génération automatique de système de dialogue construite à partir de contenus. Cette approche utilise un moteur de

recherche auquel on a ajouté des fonctionnalités de dialogue : à chaque tour, le système propose trois mots-clefs de manière à optimiser l'espérance de gain d'information.

Kawâkib est un outil assurant le feed-back entre corpus arabe et grammaire. Ce logiciel interactif en ligne démontre le bien fondé de la méthode de variation des grammaires arabes pour l'obtention de l'algorithme optimal tant au niveau de l'analyse morphologique, cruciale étant donnée la structure du système sémitique, que syntaxique ou dans le domaine de la recherche de critères pertinents et discriminants pour le filtrage des textes.

La canal chat permet aux entreprises de transformer leur site web en un véritable lieu d'achat et de service. OWI a développé un outil d'assistance aux conversations en ligne (OWI.Chat), qui analyse les messages des internautes et conseille les conseillers en temps réel.

Cet article présente ZOMBILINGO un jeu ayant un but (Game with a purpose) permettant d'annoter des corpus en syntaxe de dépendances. Les annotations créées sont librement disponibles sur le site du jeu.

Proxem édite Ubiq, une plateforme de collecte de documents et d'analyse sémantique, capable d'extraire des informations pertinentes à partir du contenu de vastes corpus. Les documents analysés sont d'une grande diversité : opinions collectées sur des sites web, emails de réclamation ou de demande d'information, réponse à des questions ouvertes dans des sondages, offres ou demandes d'emploi, etc. La reconnaissance des entités nommées joue un rôle central car c'est un préalable à d'autres traitements sémantiques. La conception d'un module de reconnaissance d'entités nommées nécessite généralement un investissement important en amont, avec une adaptation de domaine. Ubiq propose une approche d'apprentissage faiblement supervisé de l'extraction d'entités nommées qui tient compte du corpus collecté et de ressources externes

(Wikipédia). La méthode et l'outillage développés permettent de déterminer à la volée, en interaction avec l'utilisateur, la granularité des types d'entités adaptée à un corpus de texte tout-venant.

Nous proposons dans cette démonstration de présenter le logiciel Zodiac, permettant l'insertion automatique de diacritiques (accents, cédilles, etc.) dans un texte français. Zodiac prend la forme d'un complément Microsoft Word sous Windows permettant des corrections automatiques du texte au cours de la frappe. Sous Linux et Mac OS X, il est implémenté comme un programme sur ligne de commande, se prêtant naturellement à lire ses entrées sur un « pipeline » et écrire ses sorties sur la sortie standard. Implémenté en UTF-8, il met en oeuvre diverses librairies C++ utiles à certaines tâches du TAL, incluant la manipulation de modèles de langue statistiques.

Le projet STAM aborde la problématique de la transcription automatique du langage texto (SMS) et plus particulièrement la traduction des messages écrits en arabe dialectal. L'objectif du système STAM est de traduire automatiquement des textes rédigés en langage SMS dans un dialecte parlé dans le monde arabe (langue source) en un texte facilement interprétable, compréhensible et en bon français (langue cible).

Les stratégies de dialogue incrémentales offrent une meilleure réactivité, une expérience utilisateur plus aboutie et une réduction du risque de désynchronisation. Cependant, les systèmes de dialogue incrémentaux sont basés sur une architecture logicielle dont l'implantation est longue, difficile et donc coûteuse. Pour faciliter cette évolution d'architecture, nous proposons de simuler un comportement incrémental en ajoutant une surcouche à un service de dialogue traditionnel existant. DictaNum est un démonstrateur de dialogue incrémental mettant en oeuvre cette démarche. Sa tâche consiste à recueillir des numéros auprès des utilisateurs. Grâce à son fonctionnement incrémental, il autorise une correction rapide des erreurs au fil de la dictée.

Dans cet article de démonstration, nous introduisons un logiciel permettant de construire des tables de traduction de manière beaucoup plus rapide que ne le font les techniques à l'état de l'art. Cette accélération notable est obtenue par le biais d'un double échantillonnage : l'un permet la sélection d'un nombre limité de bi-phrases contenant les segments à traduire, l'autre réalise un alignement à la volée de ces bi-phrases pour extraire des exemples de traduction.

Nous proposons la démonstration d'un assistant personnel basé sur une architecture distribuée. Un portail vocal relie l'utilisateur à des applications. Celles-ci sont installées par l'utilisateur qui compose de ce fait son propre assistant personnel selon ses besoins.

Dans cet article, nous comparons la structure topologique des réseaux lexicaux avec une méthode fondée sur des marches aléatoires. Au lieu de caractériser les paires de sommets selon un critère binaire de connectivité, nous mesurons leur proximité structurelle par la probabilité relative d'atteindre un sommet depuis l'autre par une courte marche aléatoire. Parce que cette proximité rapproche les sommets d'une même zone dense en arêtes, elle permet de comparer la structure topologique des réseaux lexicaux.

Nous proposons une démonstration d'un reconnaisseur d'entités nommées du Français appris automatiquement sur le French TreeBank annoté en entités nommées.

Les méthodes de détection automatique de l'opinion dans des textes s'appuient sur l'association d'une polarité d'opinion aux mots des textes, par lexique ou par apprentissage. Or, certains mots ont des polarités qui peuvent varier selon le domaine thématique du texte. Nous proposons dans cet article une étude des mots ou groupes de mots marqueurs d'opinion au niveau du texte et qui ont une polarité changeante en fonction du domaine. Les expériences, effectuées à la fois sur des corpus français et anglais, montrent que la prise en compte de ces marqueurs permet d'améliorer

de manière significative la classification de l'opinion au niveau du texte lors de l'adaptation d'un domaine source à un domaine cible. Nous montrons également que ces marqueurs peuvent être utiles, de manière limitée, lorsque l'on est en présence d'un mélange de domaines. Si les domaines ne sont pas explicites, utiliser une séparation automatique des documents permet d'obtenir les mêmes améliorations.

Les termes-clés sont les mots ou les expressions polylexicales qui représentent le contenu principal d'un document. Ils sont utiles pour diverses applications, telles que l'indexation automatique ou le résumé automatique, mais ne sont pas toujours disponibles. De ce fait, nous nous intéressons à l'extraction automatique de termes-clés et, plus particulièrement, à la difficulté de cette tâche lors du traitement de documents appartenant à certaines disciplines scientifiques. Au moyen de cinq corpus représentant cinq disciplines différentes (archéologie, linguistique, sciences de l'information, psychologie et chimie), nous déduisons une échelle de difficulté disciplinaire et analysons les facteurs qui influent sur cette difficulté.

Les systèmes d'extraction d'information doivent faire face depuis toujours à une double difficulté : d'une part, ils souffrent d'une dépendance forte vis-à-vis du domaine pour lesquels ils ont été développés ; d'autre part, leur coût de développement pour un domaine donné est important. Le travail que nous présentons dans cet article se focalise sur la seconde problématique en proposant néanmoins une solution en relation avec la première. Plus précisément, il aborde la tâche d'étiquetage en rôles événementiels dans le cadre du remplissage de formulaire (template filling) en proposant pour ce faire de s'appuyer sur un modèle de représentation distribuée de type neuronal. Ce modèle est appris à partir d'un corpus représentatif du domaine considéré sans nécessiter en amont l'utilisation de pré-traitements linguistiques élaborés. Il fournit un espace de représentation permettant à un classifieur supervisé traditionnel de se dispenser de l'utilisation de traits complexes et variés (traits morpho-syntaxiques, syntaxiques ou sémantiques). Par une série

d'expérimentations menées sur le corpus de la campagne d'évaluation MUC-4, nous montrons en particulier que cette approche permet de dépasser les performances de l'état de l'art et que cette différence est d'autant plus importante que la taille du corpus d'entraînement est faible. Nous montrons également l'intérêt de l'adaptation de ce type de modèle au domaine traité par rapport à l'utilisation de représentations distribuées à usage générique.

D'après la sémantique des cadres de Fillmore, les mots prennent leur sens par rapport au contexte événementiel ou situationnel dans lequel ils s'inscrivent. FrameNet, une ressource lexicale pour l'anglais, définit environ 1000 cadres conceptuels couvrant l'essentiel des contextes possibles. Dans un cadre conceptuel, un prédicat appelle des arguments pour remplir les différents rôles sémantiques associés au cadre. Nous cherchons à annoter automatiquement ces rôles sémantiques, étant donné le cadre sémantique et le prédicat, à l'aide de modèles à maximum d'entropie. Nous montrons que l'utilisation de représentations distribuées de mots pour situer sémantiquement les arguments apporte une information complémentaire au modèle, et améliore notamment l'étiquetage de cadres avec peu d'exemples d'entrainement

Nous abordons la question du transfert d'annotations sémantiques, et plus spécifiquement d'étiquettes sur les prédicats, d'une langue à l'autre sur la base de corpus parallèles. Des travaux antérieurs ont transféré ces annotations directement au niveau des tokens, conduisant à un faible rappel. Nous présentons une approche globale de transfert qui agrège des informations repérées dans l'ensemble du corpus parallèle. Nous montrons que la performance de la méthode globale est supérieure aux résultats antérieurs en termes de rappel sans trop affecter la précision.

Les outils TAL statistiques robustes, et en particulier les étiqueteurs morpho-syntaxiques, utilisent souvent des descripteurs "pauvres", qui peuvent être appliqués facilement à n'importe quelle langue, mais qui ne regarde pas plus loin que 1 ou 2 tokens à droite et à gauche et ne prennent pas

en compte des classes d'équivalence syntaxiques. Bien que l'étiquetage morpho-syntaxique atteint des niveaux élevés d'exactitude (autour de 97 %), les 3 % d'erreurs qui subsistent induisent systématiquement une baisse de 3 % dans l'exactitude du parseur. Parmi les phénomènes les plus faciles à cibler à l'aide de l'injection de connaissances linguistiques plus riches sont les mots fonctionnels ambigus, tels que le mot "que" en français. Dans cette étude, nous cherchons à améliorer l'étiquetage morpho-syntaxique de "que" par l'utilisation de descripteurs ciblés et riches lors de l'entraînement, et par l'utilisation de règles symboliques qui contournent le modèle statistique lors de l'analyse. Nous atteignons une réduction du taux d'erreur de 45 % par les descripteurs riches, et de 55 % si on ajoute des règles.

Nous présentons DYALOG-SR, un analyseur syntaxique statistique par dépendances développé dans le cadre de la tâche SPRML 2013 portant sur un jeu de 9 langues très différentes. L'analyseur DYALOG-SR implémente un algorithme d'analyse par transition (à la MALT), étendu par utilisation de faisceaux et de techniques de programmation dynamique. Une des particularité de DYALOG-SR provient de sa capacité à prendre en entrée des treillis de mots, particularité utilisée lors de SPMRL13 pour traiter des treillis en Hébreu et reprise plus récemment sur des treillis produits par SXPIPE pour le français. Disposant par ailleurs avec FRMG d'un analyseur alternatif pour le français, nous avons expérimenté un couplage avec DYALOG-SR, nous permettant ainsi d'obtenir les meilleurs résultats obtenus à ce jour sur le French TreeBank.

Nous présentons une approche de la construction manuelle des ressources lexicales à large couverture fondée sur le recours à un type particulier de réseau lexical appelé système lexical. En nous appuyant sur l'expérience acquise dans le cadre de la construction du Réseau Lexical du Français (RL-fr), nous offrons tout d'abord une caractérisation formelle des systèmes lexicaux en tant que graphes d'unités lexicales de type « petits mondes » principalement organisés à partir du système des fonctions lexicales Sens-Texte. Nous apportons ensuite des arguments pour justifier la

pertinence du modèle proposé, tant du point de vue théorique qu'applicatif.

Analyser la complexité lexicale est une tâche qui, depuis toujours, a principalement retenu l'attention de psycholinguistes et d'enseignants de langues. Plus récemment, cette problématique a fait l'objet d'un intérêt grandissant dans le domaine du traitement automatique des langues (TAL) et, en particulier, en simplification automatique de textes. L'objectif de cette tâche est d'identifier des termes et des structures difficiles à comprendre par un public cible et de proposer des outils de simplification automatisée de ces contenus. Cet article aborde la question lexicale en identifiant un ensemble de prédicteurs de la complexité lexicale et en évaluant leur efficacité via une analyse corrélationnelle. Les meilleures de ces variables ont été intégrées dans un modèle capable de prédire la difficulté lexicale dans un contexte d'apprentissage du français.

Les ontologies spécifiques à un domaine ont une valeur inestimable malgré les nombreux défis liés à leur développement. Dans la plupart des cas, les bases de connaissances spécifiques à un domaine sont construites avec une portée limitée. En effet, elles ne prennent pas en compte les avantages qu'il pourrait y avoir à combiner une ontologie de spécialité à une ontologie générale. En outre, la plupart des ressources existantes manque de méta-informations sur les annotations (informations fréquentielles : de fréquent à rare ; ou des informations de pertinence : pertinent, non pertinent et inférable). Nous présentons dans cet article un réseau lexical dédié à la radiologie construit sur un réseau lexical généraliste (JeuxDeMots). Ce réseau combine poids et annotations sur des relations typées entre des termes et des concepts, un mécanisme d'inférence et de réconciliation dans le but d'améliorer la qualité et la couverture du réseau. Nous étendons ce mécanisme afin de prendre en compte non seulement les relations mais aussi les annotations. Nous décrivons la manière de laquelle les annotations améliorent le réseau en imposant de nouvelles contraintes spécialement celles basées sur la connaissance médicale. Nous présentons par la suite des résultats préliminaires.

Cet article décrit des travaux réalisés dans le cadre du développement du correcteur automatique d'un logiciel commercial d'aide à la rédaction du français. Nous voulons corriger des erreurs uniquement détectables lorsque l'antécédent de certains pronoms est connu. Nous décrivons un algorithme de résolution des anaphores pronominales intra- et inter-phrastiques s'appuyant peu sur la correspondance de la morphologie, puisque celle-ci est possiblement erronée, mais plutôt sur des informations robustes comme l'analyse syntaxique fine et des co-occurrences fiables. Nous donnons un aperçu de nos résultats sur un vaste corpus de textes réels et, tout en tentant de préciser les critères décisifs, nous montrons que certains types de corrections anaphoriques sont d'une précision respectable.

Dans cet article, nous testons deux approches distinctes pour chunker un corpus oral transcrit, en cherchant à minimiser les étapes de correction manuelle. Nous ré-utilisons tout d'abord un chunker appris sur des données écrites, puis nous tentons de ré-apprendre un chunker spécifique de l'oral à partir de données annotées et corrigées manuellement, mais en faible quantité. L'objectif est d'atteindre les meilleurs résultats possibles pour le chunker en se passant autant que possible de la correction manuelle des étiquettes POS. Nos expériences montrent qu'il est possible d'apprendre un nouveau chunker performant pour l'oral à partir d'un corpus de référence annoté de petite taille, sans intervention sur les étiquettes POS.

Dans cet article, nous proposons et évaluons un système permettant d'améliorer la qualité d'un texte bruité notamment par des erreurs orthographiques. Ce système a vocation à être intégré à une architecture complète d'extraction d'information, et a pour objectif d'améliorer les résultats d'une telle tâche. Pour chaque mot qui est inconnu d'un lexique de référence et qui n'est ni une entité nommée ni une création lexicale, notre système cherche à proposer une ou plusieurs normalisations possibles (une normalisation valide étant un mot connu dont le lemme est le même que celui de la

forme orthographiquement correcte). Pour ce faire, ce système utilise des techniques de correction automatique lexicale par règle qui reposent sur un système d'induction de règles par analogie.

Au cours des deux dernières décennies, de nombreux algorithmes ont été développés pour capturer la sémantique des mots simples en regardant leur répartition dans un grand corpus, et en comparant ces distributions dans un modèle d'espace vectoriel. En revanche, il n'est pas trivial de combiner les objets algébriques de la sémantique distributionnelle pour arriver à une dérivation d'un contenu pour des expressions complexes, composées de plusieurs mots. Notre contribution a deux buts. Le premier est d'établir une large base de comparaison pour les méthodes de composition pour le cas adjectif_nom. Cette base nous permet d'évaluer en profondeur la performance des différentes méthodes de composition. Notre second but est la proposition d'une nouvelle méthode de composition, qui est une généralisation de la méthode de Baroni & Zamparelli (2010). La performance de notre nouvelle méthode est également évaluée sur notre nouveau ensemble de test.

G-TAG est un formalisme dédié à la génération de textes. Il s'appuie sur les Grammaires d'Arbres Adjoints (TAG) qu'il étend avec des notions propres permettant de construire une forme de surface à partir d'une représentation conceptuelle. Cette représentation conceptuelle est indépendante de la langue, et le formalisme G-TAG a été conçu pour la mise en oeuvre de la synthèse dans une langue cible à partir de cette représentation. L'objectif de cet article est d'étudier G-TAG et les notions propres que ce formalisme introduit par le biais des Grammaires Catégorielles Abstraites (ACG) en exploitant leurs propriétés de réversibilité intrinsèque et leur propriété d'encodage des TAG. Nous montrons que les notions clefs d'arbre de g-dérivation et de lexicalisation en G-TAG s'expriment naturellement en ACG. La construction des formes de surface peut alors utiliser les algorithmes généraux associés aux ACG et certaines constructions absentes de G-TAG peuvent être prises en compte sans modification supplémentaire.

Les méthodes de transfert cross-lingue permettent partiellement de pallier l'absence de corpus annotés, en particulier dans le cas de langues peu dotées en ressources linguistiques. Le transfert d'étiquettes morpho-syntaxiques depuis une langue riche en ressources, complété et corrigé par un dictionnaire associant à chaque mot un ensemble d'étiquettes autorisées, ne fournit cependant qu'une information de supervision incomplète. Dans ce travail, nous reformulons ce problème dans le cadre de l'apprentissage ambigu et proposons une nouvelle méthode pour apprendre un analyseur de manière faiblement supervisée à partir d'un modèle à base d'historique. L'évaluation de cette approche montre une amélioration sensible des performances par rapport aux méthodes de l'état de l'art pour trois langues sur quatre considérées, avec des gains jusqu'à 3,9% absolus ou 35,8% relatifs.

Motivé par la problématique de construction automatique d'un corpus annoté morpho-syntaxiquement distinct d'un corpus source, nous proposons une définition générale et opérationnelle de la relation de la comparabilité entre des corpus monolingues annotés. Cette définition se veut indépendante du domaine applicatif. Nous proposons une mesure de la relation de comparabilité et une procédure de construction d'un corpus comparable. Enfin nous étudions la possibilité d'utiliser la mesure de la perplexité définie dans la théorie de l'information comme moyen de prioriser les phrases à sélectionner pour construire un corpus comparable. Nous montrons que cette mesure joue un rôle mais qu'elle n'est pas suffisante.

Le dialogue incrémental est au coeur de la recherche actuelle dans le domaine des systèmes de dialogue. Plusieurs architectures et modèles ont été publiés comme (Allen et al., 2001; Schlangen & Skantze, 2011). Ces approches ont permis de comprendre différentes facettes du dialogue incrémental, cependant, les implémenter nécessite de repartir de zéro car elles sont fondamentalement différentes des architectures qui existent dans les systèmes de dialogue actuels.

Notre approche se démarque par sa réutilisation de l'existant pour tendre vers une nouvelle génération de systèmes de dialogue qui ont un comportement incrémental mais dont le fonctionnement interne est basé sur les principes du dialogue traditionnel. Ce papier propose d'intercaler un module, appelé Scheduler, entre le service et le client. Ce Scheduler se charge de la gestion des événements asynchrones, de manière à reproduire le comportement des systèmes incrémentaux vu du client. Le service, de son côté, ne se comporte pas de manière incrémentale.

Démonette est une base de données lexicale pour le français dont les sommets (entrées lexicales) et les arcs (relations morphologiques entre les sommets) sont annotés au moyen d'informations morpho-sémantiques. Elle résulte d'une conception originale intégrant deux approches radicalement opposées : Morphonette, une ressource basée sur les analogies dérivationnelles, et DériF, un analyseur à base de règles linguistiques. Pour autant, Démonette n'est pas la simple fusion de deux ressources pré-existantes : cette base possède une architecture compatible avec l'approche lexématique de la morphologie ; son contenu peut être étendu au moyen de données issues de sources diverses. L'article présente le modèle Démonette et le contenu de sa version actuelle : 31 204 verbes, noms d'action, noms d'agent, et adjectifs de propriété dont les liens morphologiques donnent à voir des définitions bi-orientées entre ascendants et entre lexèmes en relation indirecte. Nous proposons enfin une évaluation de Démonette qui comparée à Verbaction obtient un score de 84% en rappel et de 90% en précision.

Dans cet article, nous abordons le problème de construction et d'amélioration de thésaurus distributionnels. Nous montrons d'une part que les outils de recherche d'information peuvent être directement utilisés pour la construction de ces thésaurus, en offrant des performances comparables à l'état de l'art. Nous nous intéressons d'autre part plus spécifiquement à l'amélioration des thésaurus obtenus, vus comme des graphes de plus proches voisins. En tirant parti de certaines des informations de voisinage contenues dans ces graphes nous proposons plusieurs

contributions. 1) Nous montrons comment améliorer globalement les listes de voisins en prenant en compte la réciprocité de la relation de voisinage, c'est-à-dire le fait qu'un mot soit un voisin proche d'un autre et vice-versa. 2) Nous proposons également une méthode permettant d'associer à chaque liste de voisins (i.e. à chaque entrées du thésaurus construit) un score de confiance. 3) Enfin, nous montrons comment utiliser ce score de confiance pour réordonner les listes de voisins les plus proches. Ces différentes contributions sont validées expérimentalement et offrent des améliorations significatives sur l'état de l'art.

Les modèles d'espace vectoriels mettant en oeuvre l'analyse distributionnelle s'appuient sur la redondance d'informations se trouvant dans le contexte des mots à associer. Cependant, ces modèles souffrent du nombre de dimensions considérable et de la dispersion des données dans la matrice des vecteurs de contexte. Il s'agit d'un enjeu majeur sur les corpus de spécialité pour lesquels la taille est beaucoup plus petite et les informations contextuelles moins redondantes. Nous nous intéressons au problème de la limitation de la dispersion des données sur des corpus de spécialité et proposons une méthode permettant de densifier la matrice en généralisant les contextes distributionnels. L'évaluation de la méthode sur un corpus médical en français montre qu'avec une petite fenêtre graphique et l'indice de Jaccard, la généralisation des contextes avec des relations fournies par des patrons lexico-syntaxiques permet d'améliorer les résultats, alors qu'avec une large fenêtre et le cosinus, il est préférable de généraliser avec des relations obtenues par inclusion lexicale.

Nous présentons une base de connaissances comportant des triplets de paires de verbes associés avec une relation sémantique/discursive, extraits du corpus français frWaC par une méthode s'appuyant sur la présence d'un connecteur discursif reliant deux verbes. Nous détaillons plusieurs mesures visant à évaluer la pertinence des triplets et la force d'association entre la relation sémantique/discursive et la paire de verbes. L'évaluation intrinsèque est réalisée par rapport à des

annotations manuelles. Une évaluation de la couverture de la ressource est également réalisée par rapport au corpus Annodis annoté discursivement. Cette étude produit des résultats prometteurs démontrant l'utilité potentielle de notre ressource pour les tâches d'analyse discursive mais aussi des tâches de nature sémantique.

Alors que l'importance des modèles neuronaux dans le domaine du traitement automatique des langues ne cesse de croître, les difficultés de leur apprentissage continue de freiner leur diffusion au sein de la communauté. Cet article étudie plusieurs stratégies, dont deux sont originales, pour estimer des modèles de langue neuronaux, en se focalisant sur l'ajustement du pas d'apprentissage. Les résultats expérimentaux montrent, d'une part, l'importance que revêt la conception de cette stratégie. D'autre part, le choix d'une stratégie appropriée permet d'apprendre efficacement des modèles de langue donnant lieu à des résultats à l'état de l'art en traduction automatique, avec un temps de calcul réduit et une faible influence des hyper-paramètres.

Les lexiques bilingues jouent un rôle important en recherche d'information interlingue et en traduction automatique. La construction manuelle de ces lexiques est lente et coûteuse. Les techniques d'alignement de mots sont généralement utilisées pour automatiser le processus de construction de ces lexiques à partir de corpus de textes parallèles. L'alignement de formes simples et de syntagmes nominaux à partir de corpus parallèles est une tâche relativement bien maîtrisée pour les langues à écriture latine, mais demeure une opération complexe pour l'appariement de textes n'utilisant pas la même écriture. Dans la perspective d'utiliser la translittération de noms propres de l'arabe vers l'écriture latine en alignement de mots et d'étudier son impact sur la qualité d'un lexique bilingue français-arabe construit automatiquement, cet article présente, d'une part, un système de translittération de noms propres de l'arabe vers l'écriture latine, et d'autre part, un outil d'alignement de mots simples et composés à partir de corpus de textes parallèles français-arabe. Le lexique bilingue produit par l'outil d'alignement de mots intégrant la translittération a été évalué en

utilisant deux approches : une évaluation de la qualité d'alignement à l'aide d'un alignement de référence construit manuellement et une évaluation de l'impact de ce lexique bilingue sur la qualité de traduction du système de traduction automatique statistique Moses. Les résultats obtenus montrent que la translittération améliore aussi bien la qualité de l'alignement de mots que celle de la traduction.

L'utilisation de méthodes statistiques en traduction automatique (TA) implique l'exploitation de gros corpus parallèles représentatifs de la tâche de traduction visée. La relative rareté de ces ressources fait que la question de l'adaptation au domaine est une problématique centrale en TA. Dans cet article, une étude portant sur l'adaptation thématique des données journalistiques issues d'une même source est proposée. Dans notre approche, chaque phrase d'un document est traduite avec le système de traduction approprié (c.-à-d. spécifique au thème dominant dans la phrase). Deux scénarios de traduction sont étudiés : (a) une classification manuelle, reposant sur la codification IPTC ; (b) une classification automatique. Nos expériences montrent que le scénario (b) conduit à des meilleures performances (à l'aune des métriques automatiques), que le scénario (a). L'approche la meilleure pour la métrique BLEU semble toutefois consister à ne pas réaliser d'adaptation ; on observe toutefois qu'adapter permet de lever certaines ambiguïtés sémantiques.

Nous présentons dans cet article les résultats d'expériences que nous avons menées concernant les disfluences dans le discours de patients schizophrènes (en remédiation). Ces expériences ont eu lieu dans le cadre d'une étude plus large recouvrant d'autres niveaux d'analyse linguistique, qui devraient aider à l'identification d'indices linguistiques conduisant au diagnostic de schizophrénie. Cette étude fait la part belle aux outils de traitement automatique des langues qui permettent le traitement rapide de grandes masses de données textuelles (ici, plus de 375 000 mots). La première phase de l'étude, que nous présentons ici, a confirmé la corrélation entre l'état schizophrène et le nombre de disfluences présentes dans le discours.

Notre travail porte sur la détection automatique de la reformulation paraphrastique dans les corpus oraux. L'approche proposée est une approche syntagmatique qui tient compte des marqueurs de reformulation paraphrastique et des spécificités de l'oral. L'annotation manuelle effectuée par deux annotateurs permet d'obtenir une description fine et multidimensionnelle des données de référence. Une méthode automatique est proposée afin de décider si les tours de parole comportent ou ne comportent pas des reformulations paraphrastiques. Les résultats obtenus montrent jusqu'à 66,4 % de précision. L'analyse de l'annotation manuelle indique qu'il existe peu de segments paraphrastiques avec des modifications morphologiques (flexion, dérivation ou composition) ou de segments qui montrent l'équivalence syntaxique.

La désambiguïsation morphologique d'un mot arabe consiste à identifier l'analyse morphologique appropriée correspondante à ce mot. Dans cet article, nous présentons trois modèles de désambiguïsation morphologique de textes arabes non voyellés basés sur la classification possibiliste. Cette approche traite les données imprécises dans les phases d'apprentissage et de test, étant donné que notre modèle apprend à partir de données non étiquetés. Nous testons notre approche sur deux corpus, à savoir le corpus du Hadith et le Treebank Arabe. Ces corpus contiennent des données de types différents classiques et modernes. Nous comparons nos modèles avec des classifieurs probabilistes et statistiques. Pour ce faire, nous transformons la structure des ensembles d'apprentissage et de test pour remédier au problème d'imperfection des données.

On propose un algorithme original d'analyse syntaxique déterministe en constituants pour le langage naturel inspiré de LR (Knuth, 1965). L'algorithme s'appuie sur un modèle d'apprentissage discriminant pour réaliser la désambiguisation (Collins, 2002). On montre que le modèle discriminant permet de capturer plus finement de l'information morphologique présente dans les

données, ce qui lui permet d'obtenir des résultats état de l'art en temps comme en exactitude pour l'analyse syntaxique du français.

La compréhension d'un texte s'opère à travers les niveaux d'information visuelle, logique et discursive, et leurs relations d'interdépendance. La majorité des travaux ayant étudié ces relations a été menée dans le cadre de la génération de textes, où les propriétés visuelles sont inférées à partir des éléments logiques et discursifs. Les travaux présentés ici adoptent une démarche inverse en proposant de générer automatiquement la structure organisationnelle du texte (structure logique) à partir de sa forme visuelle. Le principe consiste à (i) labelliser des blocs visuels par apprentissage afin d'obtenir des unités logiques et (ii) relier ces unités par des relations de coordination ou de subordination pour construire un arbre. Pour ces deux tâches, des Champs Aléatoires Conditionnels et un Maximum d'Entropie sont respectivement utilisés. Après apprentissage, les résultats aboutissent à une exactitude de 80,46% pour la labellisation et 97,23% pour la construction de l'arbre.

La robustesse de l'analyse probabiliste s'obtient généralement au détriment du jugement de grammaticalité sur la phrase analysée. Les analyseurs comme le Stanford Parser, ou les Re-ranking Parsers ne sont, en effet, pas capables de dissocier une analyse probable grammaticale d'une analyse probable erronée, et ce qu'elle porte sur une phrase elle-même grammaticale ou non. Dans cet article nous montrons que l'adoption d'une représentation syntaxique basée sur la théorie logique des modèles, accompagnée d'une structure syntaxique classique (par exemple de type syntagmatique), est de nature à permettre la résolution exacte de différents problèmes tels que celui du jugement de grammaticalité. Afin de démontrer la praticité et l'utilité d'une alliance entre symbolique et stochastique, nous nous appuyons sur une représentation de la syntaxe par modèles, ainsi que sur une grammaire de corpus, pour présenter une méthode de résolution exacte pour le jugement de grammaticalité d'un arbre syntagmatique probable. Nous présentons des résultats

expérimentaux sur des arbres issus d'un analyseur probabiliste, qui corroborent l'intérêt d'une telle alliance.

Nos travaux se focalisent sur la validation d'occurrences de candidats termes en contexte. Les contextes d'occurrences proviennent d'articles scientifiques des sciences du langage issus du corpus SCIENTEXT1. Les candidats termes sont identifiés par l'extracteur automatique de termes de la plate-forme TTC-TermSuite et sont ensuite projetés dans les textes. La problématique générale de cet article est d'étudier dans quelle mesure les contextes sont à même de fournir des critères linguistiques pertinents pour valider ou rejeter chaque occurrence de candidat terme selon qu'elle relève d'un usage terminologique en sciences du langage ou non (langue générale, transdisciplinaire, autre domaine scientifique). Pour répondre à cette question, nous comparons deux méthodes d'exploitation (l'une inspirée de la textométrie et l'autre de Lesk) avec des contextes d'occurrences du même corpus annotés manuellement et mesurons si une annotation sémantique des contextes améliore l'exactitude des choix réalisés automatiquement.

Dans cet article, nous nous intéressons aux noms sous-spécifiés, qui forment une classe d'indices de l'organisation discursive. Ces indices ont été peu étudiés dans le cadre de l'analyse du discours et en traitement automatique des langues. L'objectif est d'effectuer une étude linguistique de leur participation à la structuration discursive, notamment lorsqu'ils interviennent dans des séquences organisationnelles fréquentes (e.g. le patron Problème-Solution). Dans cet article, nous présentons les différentes étapes mises en oeuvre pour identifier automatiquement ces noms en corpus. En premier lieu, nous détaillons la construction d'un lexique de noms sous-spécifiés pour le français à partir d'un corpus constitué de 7 années du journal Le Monde. Puis nous montrons comment utiliser des techniques fondées sur la fouille de données séquentielles pour acquérir de nouvelles constructions syntaxiques caractéristiques des emplois de noms sousspécifiés. Enfin, nous présentons une méthode d'identification automatique des occurrences de noms sous-spécifiés et

son évaluation.

Les bases de données sont de plus en plus courantes et prennent de plus en plus d'ampleur au sein des applications et sites Web actuels. Elles sont souvent amenées à être utilisées par des personnes n'ayant pas une grande compétence en la matière et ne connaissant pas rigoureusement leur structure. C'est pour cette raison que des traducteurs du langage naturel aux requêtes SQL sont développés. Malheureusement, la plupart de ces traducteurs se cantonnent à une seule base du fait de la spécificité de l'architecture de celle-ci. Dans cet article, nous proposons une méthode visant à pouvoir interroger n'importe quelle base de données à partir de questions en français. Nous évaluons notre application sur deux bases à la structure différente et nous montrons également qu'elle supporte plus d'opérations que la plupart des autres traducteurs.

La désambiguïsation lexicale permet d'améliorer de nombreuses applications en traitement automatique des langues (TAL) comme la recherche d'information, l'extraction d'information, la traduction automatique, ou la simplification lexicale de textes. Schématiquement, il s'agit de choisir quel est le sens le plus approprié pour chaque mot d'un texte. Une des approches classiques consiste à estimer la similarité sémantique qui existe entre les sens de deux mots puis de l'étendre à l'ensemble des mots du texte. La méthode la plus directe donne un score de similarité à toutes les paires de sens de mots puis choisit la chaîne de sens qui retourne le meilleur score (on imagine la complexité exponentielle liée à cette approche exhaustive). Dans cet article, nous proposons d'utiliser une méta-heuristique d'optimisation combinatoire qui consiste à choisir les voisins les plus proches par sélection distributionnelle autour du mot à désambiguïser. Le test et l'évaluation de notre méthode portent sur un corpus écrit en langue française en se servant du réseau sémantique BabelNet. Le taux d'exactitude obtenu est de 78% sur l'ensemble des noms et des verbes choisis pour l'évaluation.

Cet article aborde la question de la détection des expressions d'attitude ? i.e affect, d'appréciation et de jugement (Martin & White, 2005) ? dans le contenu verbal de l'utilisateur au cours d'interactions en face-à-face avec un agent conversationnel animé. Il propose un positionnement en termes de modèles et de méthodes pour le développement d'un système de détection adapté aux buts communicationnels de l'agent et à une parole conversationnelle. Après une description du modèle théorique de référence choisi, l'article propose un modèle d'annotation des attitudes dédié l'exploration de ce phénomène dans un corpus d'interaction humain-agent. Il présente ensuite une première version de notre système. Cette première version se concentre sur la détection des expressions d'attitudes pouvant référer à ce qu'aime ou n'aime pas l'utilisateur. Le système est conçu selon une approche symbolique fondée sur un ensemble de règles sémantiques et de représentations logico-sémantiques des énoncés.

Les travaux menés dans le cadre du résumé automatique de texte ont montré des résultats à la fois très encourageants mais qui sont toujours à améliorer. La problématique du résumé automatique ne cesse d'évoluer avec les nouveaux champs d'application qui s'imposent, ce qui augmente les contraintes liées à cette tâche. Nous nous inté- ressons au résumé extractif multi-document dynamique. Pour cela, nous examinons les différentes approches existantes en mettant l'accent sur les travaux les plus récents. Nous montrons ensuite que la performance des systèmes de résumé multi-document et dynamique est encore modeste. Trois contraintes supplémentaires sont ajoutées : la redondance inter-document, la redondance à travers le temps et la grande taille des données à traiter. Nous essayons de déceler les insuffisances des systèmes existants afin de bien définir notre problématique et guider ainsi nos prochains travaux.

Cet article présente certaines questions de recherche liées au projet COCO. L'ambition de ce projet est de valoriser les ressources éducatives et académiques en exploitant au mieux les différents médias disponibles (vidéos de cours ou de présentations d'articles, manuels éducatifs, articles

scientifiques, présentations, etc). Dans un premier temps, nous décrirons le problème d'utilisation jointe de ressources multimédias éducatives ou scientifiques pour ensuite introduire l'état de l'art dans les domaines concernés. Cela nous permettra de présenter quelques questions de recherche sur lesquelles porteront des études ultérieures. Enfin nous finirons en introduisant trois prototypes développés pour analyser ces questions.

Le développement du Web 2.0 et le processus de création et de consommation massive de contenus générés par les utilisateurs qu'elle a enclenché a permis le développement de nouveaux types d'interactions chez les internautes. En particulier, nous nous intéressons au développement du support en ligne et des plate-formes d'entraide. En effet, les archives de conversations en ligne porteuses de demandes d'assistance représentent une ressource inestimable, mais peu exploitée. L'exploitation de cette ressource permettrait non seulement d'améliorer les systèmes liés à la résolution collaborative des problèmes, mais également de perfectionner les canaux de support proposés par les entreprises opérant sur le web. Pour ce faire, il est cependant nécessaire de définir un cadre formel pour l'analyse discursive de ce type de conversations. Cet article a pour objectif de présenter l'état de la recherche en analyse des conversations écrites en ligne, sous différents médiums, et de montrer dans quelle mesure les différentes méthodes exposées dans la littérature peuvent être appliquées à des conversations fonctionnelles inscrites dans le cadre de la résolution collaborative des problèmes utilisateurs.

Marge infusé algorithmes détendus (MIRAS) dominent modèle de tuning dans la traduction automatique statistique dans le cas des grandes caractéristiques de l'échelle, mais ils sont également célèbres pour la complexité de mise en oeuvre. Nous introduisons une nouvelle méthode, qui concerne une liste des N meilleures comme une permutation et minimise la perte Plackett-Luce de permutations rez-de-vérité. Des expériences avec des caractéristiques à grande échelle démontrent que, la nouvelle méthode est plus robuste que MERT ; si ce est seulement à

rattacher avec Miras, il a un avantage comparativement, plus facile à mettre en oeuvre.

Dans cet article, nous présentons une utilisation des assistants des preuves pour traiter l'inférence en Language Naturel (NLI). D' abord, nous proposons d'utiliser les theories des types modernes comme langue dans laquelle traduire la sémantique du langage naturel. Ensuite, nous implémentons cette sémantique dans l'assistant de preuve Coq pour raisonner sur ceux-ci. En particulier, nous évaluons notre proposition sur un sous-ensemble de la suite de tests FraCas, et nous montrons que 95.2% des exemples peuvent être correctement prédits. Nous discutons ensuite la question de l'automatisation et il est démontré que le langage de tactiques de Coq permet de construire des tactiques qui peuvent automatiser entièrement les preuves, au moins pour les cas qui nous intéressent.

En analyse de discours ou d'opinion, savoir caractériser la connotation générale d'un texte, les sentiments qu'il véhicule, est une aptitude recherchée, qui suppose la constitution préalable d'une ressource lexicale de polarité. Au sein du réseau lexical JeuxDeMots, nous avons mis au point LikeIt, un jeu qui permet d'affecter une valeur positive, négative, ou neutre à un terme, et de constituer ainsi pour chaque terme, à partir des votes, une polarité résultante. Nous présentons ici l'analyse quantitative des données de polarité obtenues, ainsi que la méthode pour les valider qualitativement.

Dans cet article, nous combinons annotations manuelle et automatique pour identifier les verbes utilisés pour introduire un médicament dans les messages sur les forums de santé. Cette information est notamment utile pour identifier la relation entre un médicament et un effet secondaire. La mention d'un médicament dans un message ne garantit pas que l'utilisateur a pris ce traitement mais qu'il effectue un retour. Nous montrons ensuite que ces verbes peuvent servir pour extraire automatiquement des variantes de noms de médicaments. Nous estimons que l'analyse de

ces variantes pourrait permettre de modéliser les erreurs faites par les usagers des forums lorsqu'ils écrivent les noms de médicaments, et améliorer en conséquence les systèmes de recherche d'information.

Ce papier présente une méthode de traitement de documents parlés intégrant une représentation fondée sur un espace thématique dans un réseau de neurones artificiels (ANN) employé comme classifieur de document. La méthode proposée consiste à configurer la topologie d'un ANN ainsi que d'initialiser les connexions de celui-ci à l'aide des espaces thématiques appris précédemment. Il est attendu que l'initialisation fondée sur les probabilités thématiques permette d'optimiser le processus d'optimisation des poids du réseau ainsi qu'à accélérer la phase d'apprentissage tout en amélioration la précision de la classification d'un document de test. Cette méthode est évaluée lors d'une tâche de catégorisation de dialogues parlés entre des utilisateurs et des agents du service d'appels de la Régie Aut-onome Des Transports Parisiens (RATP). Les résultats montrent l'intérêt de la méthode proposée d'initialisation d'un réseau, avec un gain observé de plus de 4 points en termes de bonne classification comparativement à l'initialisation aléatoire. De plus, les expérimentations soulignent que les performances sont faiblement dépendantes de la topologie du ANN lorsque les poids de la couche cachée sont initialisés au moyen des espaces de thèmes issus d'une allocation latente de Dirichlet ou latent Dirichlet Allocation (LDA) en comparaison à une initialisation empirique.

Cet article présente le repérage manuel des connecteurs de discours dans le corpus FTB (French Treebank) déjà annoté pour la morpho-syntaxe. C'est la première étape de l'annotation discursive complète de ce corpus. Il s'agit de projeter sur le corpus les éléments répertoriés dans LexConn, lexique des connecteurs du français, et de filtrer les occurrences de ces éléments qui n'ont pas un emploi discursif mais par exemple un emploi d'adverbe de manière ou de préposition introduisant un complément sous-catégorisé. Plus de 10 000 connecteurs ont ainsi été repérés.

Dans cet article, nous proposons une mesure de distance entre phrases fondée sur la distance de Levenshtein, doublement pondérée par la fréquence des mots et par le type d'opération réalisée. Nous l'évaluons au sein d'un système de résumé automatique dont la méthode de calcul est volontairement limitée à une approche fondée sur la similarité entre phrases. Nous sommes donc ainsi en mesure d'évaluer indirectement la performance de cette nouvelle mesure de distance.

Dans cet article, nous nous intéressons à la classification contextuelle d'entités nommées de type « film ». Notre travail s'inscrit dans un cadre applicatif dont le but est de repérer, dans un texte, un titre de film contenu dans un catalogue (par exemple catalogue de films disponibles en VoD). Pour ce faire, nous combinons deux approches : nous partons d'un système à base de règles, qui présente une bonne précision, que nous couplons avec un modèle de langage permettant d'augmenter le rappel. La génération peu coûteuse de données d'apprentissage pour le modèle de langage à partir de Wikipedia est au coeur de ce travail. Nous montrons, à travers l'évaluation de notre système, la difficulté de classification des entités nommées de type « film » ainsi que la complémentarité des approches que nous utilisons pour cette tâche.

Nous présentons des travaux préliminaires sur une approche permettant d'ajouter des termes bilingues à un système de Traduction Automatique Statistique (TAS) à base de segments. Les termes sont non seulement inclus individuellement, mais aussi avec des contextes les englobant. Tout d'abord nous générons ces contextes en généralisant des motifs (ou patrons) observés pour des mots de même nature syntaxique dans un corpus bilingue. Enfin, nous filtrons les contextes qui n'atteignent pas un certain seuil de confiance, à l'aide d'une méthode de sélection de bi-segments inspirée d'une approche de sélection de données, précédemment appliquée à des textes bilingues alignés.

Dans cette contribution, nous présentons une étude sur la stylistique computationnelle des textes de la littérature classiques française fondée sur une approche conduite par données, où la découverte des motifs linguistiques intéressants se fait sans aucune connaissance préalable. Nous proposons une mesure objective capable de capturer et d'extraire des motifs syntaxiques stylistiques significatifs à partir d'un oeuvre d'un auteur donné. Notre hypothèse de travail est fondée sur le fait que les motifs syntaxiques les plus pertinents devraient refléter de manière significative le choix stylistique de l'auteur, et donc ils doivent présenter une sorte de comportement de sur-représentation contrôlé par les objectifs de l'auteur. Les résultats analysés montrent l'efficacité dans l'extraction de motifs syntaxiques intéressants dans le texte littéraire français classique, et semblent particulièrement prometteurs pour les analyses de ce type particulier de texte.

Dans cet article, nous proposons une méthode de résumé automatique fondée sur l'utilisation d'un algorithme génétique pour parcourir l'espace des résumés candidats couplé à un calcul de divergence de distribution de probabilités de n-grammes entre résumés candidats et documents source. Cette méthode permet de considérer un résumé non plus comme une accumulation de phrases indépendantes les unes des autres, mais comme un texte vu dans sa globalité. Nous la comparons à une des meilleures méthodes existantes fondée sur la programmation linéaire en nombre entier, et montrons son efficacité sur le corpus TAC 2009.

Le travail présenté dans cet article se situe dans le cadre de l'identification de termes spécialisés (unités de mesure) à partir de données textuelles pour enrichir une Ressource Termino-Ontologique (RTO). La première étape de notre méthode consiste à prédire la localisation des variants d'unités de mesure dans les documents. Nous avons utilisé une méthode reposant sur l'apprentissage supervisé. Cette méthode permet de réduire sensiblement l'espace de recherche des variants tout en restant dans un contexte optimal de recherche (réduction de 86% de l'espace de recherché sur le corpus étudié). La deuxième étape du processus, une fois l'espace de recherche réduit aux

variants d'unités, utilise une nouvelle mesure de similarité permettant d'identifier automatiquement les variants découverts par rapport à un terme d'unité déjà référencé dans la RTO avec un taux de précision de 82% pour un seuil au dessus de 0.6 sur le corpus étudié.

Le nettoyage de documents issus du web est une tâche importante pour le TAL en général et pour la constitution de corpus en particulier. Cette phase est peu traitée dans la littérature, pourtant elle n'est pas sans influence sur la qualité des informations extraites des corpus. Nous proposons deux types d'évaluation de cette tâche de détourage : (I) une évaluation intrinsèque fondée sur le contenu en mots, balises et caractères ; (II) une évaluation extrinsèque fondée sur la tâche, en examinant l'effet du détourage des documents sur le système placé en aval de la chaîne de traitement. Nous montrons que les résultats ne sont pas cohérents entre ces deux évaluations ainsi qu'entre les différentes langues. Ainsi, le choix d'un outil de détourage devrait être guidé par la tâche visée plutôt que par la simple évaluation intrinsèque.

La fouille d'opinion ciblée (aspect-based sentiment analysis) fait l'objet ces dernières années d'un intérêt particulier, visible dans les sujets des récentes campagnes d'évaluation comme SemEval 2014 et 2015 ou bien DEFT 2015. Cependant les corpus annotés et publiquement disponibles permettant l'évaluation de cette tâche sont rares. Dans ce travail nous présentons en premier lieu un corpus français librement accessible de 10 000 tweets manuellement annotés. Nous accompagnons ce corpus de résultats de référence pour l'extraction de marqueurs d'opinion non supervisée. Nous présentons ensuite une méthode améliorant les résultats de cette extraction, en suivant une approche semi-supervisée.

Les banques terminologiques et les dictionnaires sont des ressources précieuses qui facilitent l'accès aux connaissances des domaines spécialisés. Ces ressources sont souvent assez pauvres et ne proposent pas toujours pour un terme à illustrer des exemples permettant d'appréhender le

sens et l'usage de ce terme. Dans ce contexte, nous proposons de mettre en oeuvre la notion de Contextes Riches en Connaissances (CRC) pour extraire directement de corpus spécialisés des exemples de contextes illustrant son usage. Nous définissons un cadre unifié pour exploiter tout à la fois des patrons de connaissances et des collocations avec une qualité acceptable pour une révision humaine.

L'objectif de cet article est de présenter tout d'abord dans ses grandes lignes le projet Anamètre qui a pour objet le traitement automatique des formes métriques de la poésie et du théâtre français du début du XVIIe au début du XXe siècle. Nous présenterons ensuite un programme de calcul automatique des mètres appliqué à notre corpus dans le cadre d'une approche déterministe en nous appuyant sur la méthode métricométrique de B. de Cornulier ainsi que la procédure d'appariement des rimes et la détermination des schémas de strophes dans les suites périodiques et les formes fixes.

Cet article présente CROC 1 (Coreference Resolution for Oral Corpus), un premier système de résolution des co-références en français reposant sur des techniques d'apprentissage automatique. Une des spécificités du système réside dans son apprentissage sur des données exclusivement orales, à savoir ANCOR (anaphore et co-référence dans les corpus oraux), le premier corpus de français oral transcrit annoté en relations anaphoriques. En l'état actuel, le système CROC nécessite un repérage préalable des mentions. Nous détaillons les choix des traits ? issus du corpus ou calculés ? utilisés par l'apprentissage, et nous présentons un ensemble d'expérimentations avec ces traits. Les scores obtenus sont très proches de ceux de l'état de l'art des systèmes conçus pour l'écrit. Nous concluons alors en donnant des perspectives sur la réalisation d'un système end-to-end valable à la fois pour l'oral transcrit et l'écrit.

Les recherches autour de la désambiguïsation sémantique traitent de la question du sens à

accorder à différentes occurrences d'un mot ou plus largement d'une unité lexicale. Dans cet article, nous nous intéressons à l'ambiguïté d'un terme en domaine de spécialité. Nous posons les premiers jalons de nos recherches sur une question connexe que nous nommons le diagnostic d'ambiguïté. Cette tâche consiste à décider si une occurrence d'un terme est ou n'est pas ambiguë. Nous mettons en oeuvre une approche d'apprentissage supervisée qui exploite un corpus d'articles de sciences humaines rédigés en français dans lequel les termes ambigus ont été détectés par des experts. Le diagnostic s'appuie sur deux types de traits : syntaxiques et positionnels. Nous montrons l'intérêt de la structuration du texte pour établir le diagnostic d'ambiguïté.

Les données médicales étant de plus en plus informatisées, le traitement sémantiquement efficace des rapports médicaux est devenu une nécessité. La recherche d'images radiologiques peut être grandement facilitée grâce à l'indexation textuelle des comptes rendus associés. Nous présentons un algorithme d'augmentation d'index de comptes rendus fondé sur la propagation d'activation sur un réseau lexico-sémantique généraliste.

Cet article présente une méthode par apprentissage supervisé pour la détection de l'ironie dans les tweets en français. Un classifieur binaire utilise des traits de l'état de l'art dont les performances sont reconnues, ainsi que de nouveaux traits issus de notre étude de corpus. En particulier, nous nous sommes intéressés à la négation et aux oppositions explicites/implicites entre des expressions d'opinion ayant des polarités différentes. Les résultats obtenus sont encourageants.

Dans cet article, nous présentons une ressource linguistique, Morfetik, développée au LDI. Après avoir présenté le modèle sous-jacent et spécifié les modalités de sa construction, nous comparons cette ressource avec d'autres ressources du français : le GLAFF, le LEFF, Morphalou et Dicolecte. Nous étudions ensuite la couverture lexicale de ces dictionnaires sur trois corpus, le Wikipedia français, la version française de Wacky et les dix ans du Monde. Nous concluons par un programme

de travail permettant de mettre à jour de façon continue la ressource lexicographique du point de vue des formes linguistiques, en connectant la ressource à un corpus continu.

Dans cet article, nous présentons une modélisation de la morphologie dérivationnelle de l'arabe utilisant le cadre métagrammatical offert par XMG. Nous démontrons que l'utilisation de racines et patrons abstraits comme morphèmes atomiques sous-spécifiés offre une manière élégante de traiter l'interaction entre morphologie et sémantique.

Nous présentons ici 4-couv, un nouveau corpus arboré d'environ 3 500 phrases, constitué d'un ensemble de quatrièmes de couverture, étiqueté et analysé automatiquement puis corrigé et validé à la main. Il répond à des besoins spécifiques pour des projets de linguistique expérimentale, et vise à rester compatible avec les autres treebanks existants pour le français. Nous présentons ici le corpus lui-même ainsi que les outils utilisés pour les différentes étapes de son élaboration : choix des textes, étiquetage, parsing, correction manuelle.

Dans cet article nous proposons une première étude descriptive d'un corpus de conversations de type tchat issues d'un centre de contact d'assistance. Les dimensions lexicales, syntaxiques et interactionnelles sont analysées. L'étude parallèle de transcriptions de conversations téléphoniques issues d'un centre d'appel dans le même domaine de l'assistance permet d'établir des comparaisons entre ces deux modes d'interaction. L'analyse révèle des différences marquées en termes de déroulement de la conversation, avec une plus grande efficacité pour les conversations de type tchat malgré un plus grand étalement temporel. L'analyse lexicale et syntaxique révèle également des différences de niveaux de langage avec une plus grande proximité entre le client et le téléconseiller à l'oral que pour les tchats où le décalage entre le style adopté par le téléconseiller et l'expression du client est plus important.

Notre société développe un moteur de recherche (MR) sémantique basé sur la reformulation de requête. Notre MR s'appuie sur un lexique que nous avons construit en nous inspirant de la Théorie Sens-Texte (TST). Nous présentons ici notre ressource lexicale et indiquons comment nous l'enrichissons et la maintenons en fonction des besoins détectés à l'usage. Nous abordons également la question de l'adaptation de la TST à nos besoins.

Les travaux présentés portent sur l'extraction automatique d'unités sémantiques et l'évaluation de leur pertinence pour des conversations téléphoniques. Le corpus utilisé est le corpus français DECODA. L'objectif de la tâche est de permettre l'étiquetage automatique en thème de chaque conversation. Compte tenu du caractère spontané de ce type de conversations et de la taille du corpus, nous proposons de recourir à une stratégie semi-supervisée fondée sur la construction d'une ontologie et d'un apprentissage actif simple : un annotateur humain analyse non seulement les listes d'unités sémantiques candidates menant au thème mais étudie également une petite quantité de conversations. La pertinence de la relation unissant les unités sémantiques conservées, le sous-thème issu de l'ontologie et le thème annoté est évaluée par un DNN, prenant en compte une représentation vectorielle du document. L'intégration des unités sémantiques retenues dans le processus de classification en thème améliore les performances.

L'étiquetage morpho-syntaxique est une tâche fondamentale du Traitement Automatique de la Langue, sur laquelle reposent souvent des traitements plus complexes tels que l'extraction d'information ou la traduction automatique. L'étiquetage en domaine de spécialité est limité par la disponibilité d'outils et de corpus annotés spécifiques au domaine. Dans cet article, nous présentons le développement d'un corpus clinique du français annoté morpho-syntaxiquement à l'aide d'un jeu d'étiquettes issus des guides d'annotation French Treebank et Multitag. L'analyse de ce corpus nous permet de caractériser le domaine clinique et de dégager les points clés pour l'adaptation d'outils d'analyse morpho-syntaxique à ce domaine. Nous montrons également les limites d'un outil

entraîné sur un corpus journalistique appliqué au domaine clinique. En perspective de ce travail, nous envisageons une application du corpus clinique annoté pour améliorer l'étiquetage morpho-syntaxique des documents cliniques en français.

Nous présentons une typologie de liens pour un corpus multimédia ancré dans le domaine journalistique. Bien que plusieurs typologies aient été créées et utilisées par la communauté, aucune ne permet de répondre aux enjeux de taille et de variété soulevés par l'utilisation d'un corpus large comprenant des textes, des vidéos, ou des émissions radiophoniques. Nous proposons donc une nouvelle typologie, première étape visant à la création et la catégorisation automatique de liens entre des fragments de documents afin de proposer de nouveaux modes de navigation au sein d'un grand corpus. Plusieurs exemples d'instanciation de la typologie sont présentés afin d'illustrer son intérêt.

Dans le cadre de l'analyse en dépendances du français, le phénomène de la non-projectivité est peu pris en compte, en majeure partie car les donneés sur lesquelles sont entraînés les analyseurs représentent peu ou pas ces cas particuliers. Nous présentons, dans cet article, un nouveau corpus en dépendances pour le français, librement disponible, contenant un nombre substantiel de dépendances non-projectives. Ce corpus permettra d'étudier et de mieux prendre en compte les cas de non-projectivité dans l'analyse du français.

La construction d'outils d'analyse linguistique pour les langues faiblement dotées est limitée, entre autres, par le manque de corpus annotés. Dans cet article, nous proposons une méthode pour construire automatiquement des outils d'analyse via une projection interlingue d'annotations linguistiques en utilisant des corpus parallèles. Notre approche n'utilise pas d'autres sources d'information, ce qui la rend applicable à un large éventail de langues peu dotées. Nous proposons d'utiliser les réseaux de neurones récurrents pour projeter les annotations d'une langue à une autre

(sans utiliser d'information d'alignement des mots). Dans un premier temps, nous explorons la tâche d'annotation morpho-syntaxique. Notre méthode combinée avec une méthode de projection d'annotation basique (utilisant l'alignement mot à mot), donne des résultats comparables à ceux de l'état de l'art sur une tâche similaire.

Dans cet article, nous nous intéressons au titrage automatique des segments issus de la segmentation thématique de journaux télévisés. Nous proposons d'associer un segment à un article de presse écrite collecté le jour même de la diffusion du journal. La tâche consiste à apparier un segment à un article de presse à l'aide d'une mesure de similarité. Cette approche soulève plusieurs problèmes, comme la sélection des articles candidats, une bonne représentation du segment et des articles, le choix d'une mesure de similarité robuste aux imprécisions de la segmentation. Des expériences sont menées sur un corpus varié de journaux télévisés français collectés pendant une semaine, conjointement avec des articles aspirés à partir de la page d'accueil de Google Actualités. Nous introduisons une métrique d'évaluation reflétant la qualité de la segmentation, du titrage ainsi que la qualité conjointe de la segmentation et du titrage. L'approche donne de bonnes performances et se révèle robuste à la segmentation thématique.

Cet article présente en détails notre participation à la tâche 4 de SemEval2014 (Analyse de Sentiments associés aux Aspects). Nous présentons la tâche et décrivons précisément notre système qui consiste en une combinaison de composants linguistiques et de modules de classification. Nous exposons ensuite les résultats de son évaluation, ainsi que les résultats des meilleurs systèmes. Nous concluons par la présentation de quelques nouvelles expériences réalisées en vue de l'amélioration de ce système.

Dans le but de proposer une caractérisation des relations de discours liées à la causalité, nous avons été amenée à constituer et annoter notre propre corpus d'étude : la ressource EXPLICADIS

(EXPlication et Argumentation en DIScours). Cette ressource a été construite dans la continuité d'une ressource déjà disponible, le corpus ANNODIS. Proposant une annotation plus précise des relations causales sur un ensemble de textes diversifiés en genres textuels, EXPLICADIS est le premier corpus de ce type constitué spécifiquement pour l'étude des relations de discours causales.

En sémantique distributionnelle, le sens des mots est modélisé par des vecteurs qui représentent leur distribution en corpus. Les modèles étant souvent calculés sur des corpus sans pré-traitement linguistique poussé, ils ne permettent pas de rendre bien compte de la compositionnalité morphologique des mots-formes. Nous proposons une méthode pour décomposer les vecteurs de mots en vecteurs lexicaux et flexionnels.

La période « préclassique » du français s'étend sur tout le XVIe siècle et la première moitié du XVIIe siècle. Cet état de langue écrite, qui accompagne les débuts de l'imprimerie, est relativement proche du français moderne, mais se caractérise par une grande variabilité graphique. Il s'agit de l'un des moins bien dotés en termes de ressources. Nous présentons ici la construction d'un lexique, d'un corpus d'apprentissage et d'un modèle de langage pour la période préclassique, à partir de ressources du français moderne.

En classification de textes, la plupart des méthodes fondées sur des classifieurs statistiques utilisent des mots, ou des combinaisons de mots contigus, comme descripteurs. Si l'on veut prendre en compte plus d'informations le nombre de descripteurs non contigus augmente exponentiellement. Pour pallier à cette croissance, la fouille de motifs séquentiels permet d'extraire, de façon efficace, un nombre réduit de descripteurs qui sont à la fois fréquents et pertinents grâce à l'utilisation de contraintes. Dans ce papier, nous comparons l'utilisation de motifs fréquents sous contraintes et l'utilisation de motifs ?-libres, comme descripteurs. Nous montrons les avantages et inconvénients de chaque type de motif.

Pour concevoir des générateurs automatiques de texte génériques qui soient facilement réutilisables d'une langue et d'une application à l'autre, il faut modéliser les principaux phénomènes linguistiques qu'on retrouve dans les langues en général. Un des phénomènes fondamentaux qui demeurent problématiques pour le TAL est celui des collocations, comme grippe carabinée, peur bleue ou désir ardent, où un sens (ici, l'intensité) ne s'exprime pas de la même façon selon l'unité lexicale qu'il modifie. Dans la lexicographie explicative et combinatoire, on modélise les collocations au moyen de fonctions lexicales qui correspondent à des patrons récurrents de collocations. Par exemple, les expressions mentionnées ici se décrivent au moyen de la fonction Magn : Magn(PEUR) = BLEUE, Magn(GRIPPE) = CARABINÉE, etc. Il existe des centaines de fonctions lexicales. Dans cet article, nous nous intéressons à l'implémentation d'un sous-ensemble de fonctions qui décrivent les verbes supports et certains types de modificateurs.

Dans cet article, nous nous intéressons à la manière dont sont exprimés les liens qui existent entre un traitement médical et un effet secondaire. Parce que les patients se tournent en priorité vers internet, nous fondons cette étude sur un corpus annoté de messages issus de forums de santé en français. L'objectif de ce travail consiste à mettre en évidence des éléments linguistiques (connecteurs logiques et expressions temporelles) qui pourraient être utiles pour des systèmes automatiques de repérage des effets secondaires. Nous observons que les modalités d'écriture sur les forums ne permettent pas de se fonder sur les expressions temporelles. En revanche, les connecteurs logiques semblent utiles pour identifier les effets secondaires.

Dans cet article, nous montrons qu'un graphe à 1 plus proche voisin (graphe 1-PPV) offre différents moyens d'explorer les voisinages sémantiques captés par un modèle distributionnel. Nous vérifions si les composantes connexes de ce graphe, qui représentent des ensembles de mots apparaissant dans des contextes similaires, permettent d'identifier des ensembles d'unités lexicales qui évoquent

un même cadre sémantique. Nous illustrons également différentes façons d'exploiter le graphe 1-PPV afin d'explorer un modèle ou de comparer différents modèles.

Les représentations vectorielles continues des mots sont en plein essor et ont déjà été appliquées avec succès à de nombreuses tâches en traitement automatique de la langue (TAL). Dans cet article, nous proposons d'intégrer l'information temporelle issue du contexte des mots au sein des architectures fondées sur les sacs-de-mots continus (continuous bag-of-words ou CBOW) ou sur les Skip-Grams. Ces approches sont manipulées au travers d'un réseau de neurones, l'architecture CBOW cherchant alors à prédire un mot sachant son contexte, alors que l'architecture Skip-Gram prédit un contexte sachant un mot. Cependant, ces modèles, au travers du réseau de neurones, s'appuient sur des représentations en sac-de-mots et ne tiennent pas compte, explicitement, de l'ordre des mots. En conséquence, chaque mot a potentiellement la même influence dans le réseau de neurones. Nous proposons alors une méthode originale qui intègre l'information temporelle des contextes des mots en utilisant leur position relative. Cette méthode s'inspire des modèles contextuels continus. L'information temporelle est traitée comme coefficient de pondération, en entrée du réseau de neurones par le CBOW et dans la couche de sortie par le Skip-Gram. Les premières expériences ont été réalisées en utilisant un corpus de test mesurant la qualité de la relation sémantique-syntactique des mots. Les résultats préliminaires obtenus montrent l'apport du contexte des mots, avec des gains de 7 et 7,7 points respectivement avec l'architecture Skip-Gram et l'architecture CBOW.

Nous nous intéressons dans cet article à l'apprentissage automatique d'un étiqueteur mopho-syntaxique pour les tweets en anglais. Nous proposons tout d'abord un jeu d'étiquettes réduit avec 17 étiquettes différentes, qui permet d'obtenir de meilleures performances en exactitude par rapport au jeu d'étiquettes traditionnel qui contient 45 étiquettes. Comme nous disposons de peu de tweets étiquetés, nous essayons ensuite de compenser ce handicap en ajoutant dans

l'ensemble d'apprentissage des données issues de textes bien formés. Les modèles mixtes obtenus permettent d'améliorer les résultats par rapport aux modèles appris avec un seul corpus, qu'il soit issu de Twitter ou de textes journalistiques.

L'article présente une étude des propriétés linguistiques (lexicales, morpho-syntaxiques, syntaxiques) permettant la classification automatique de documents selon leur genre (articles scientifiques et articles de vulgarisation), dans deux domaines différentes (médecine et informatique). Notre analyse, effectuée sur des corpus comparables en genre et en thèmes disponibles en français, permet de valider certaines propriétés identifiées dans la littérature comme caractéristiques des discours académiques ou de vulgarisation scientifique. Les premières expériences de classification évaluent l'influence de ces propriétés pour l'identification automatique du genre pour le cas spécifique des textes scientifiques ou de vulgarisation.

Le présent article s'intéresse à la détection et à la désambiguïsation des comparaisons figuratives. Il décrit un algorithme qui utilise un analyseur syntaxique de surface (chunker) et des règles manuelles afin d'extraire et d'analyser les (pseudo-)comparaisons présentes dans un texte. Cet algorithme, évalué sur un corpus de textes littéraires, donne de meilleurs résultats qu'un système reposant sur une analyse syntaxique profonde.

Dans cet article, nous présentons une méthodologie pour l'identification de messages suspectés d'être produits par des Community Managers à des fins commerciales déguisées dans des documents du Web 2.0. Le champ d'application est la malbouffe (junkfood) et le corpus est multi-lingue (anglais, chinois, français). Nous exposons dans un premier temps la stratégie de constitution et d'annotation de nos corpus, en explicitant notamment notre guide d'annotation, puis nous développons la méthode adoptée, basée sur la combinaison d'une analyse textométrique et d'un apprentissage supervisé.

MEDITE est un logiciel d'alignement de textes permettant l'identification de transformations entre une version et une autre d'un même texte. Dans ce papier nous présentons les aspects théoriques et techniques de MEDITE.

Phoebus est un logiciel d'extraction de réutilisations dans des textes littéraires. Il a été développé comme un outil d'analyse littéraire assistée par ordinateur. Dans ce contexte, ce logiciel détecte automatiquement et explore des réseaux de réutilisation textuelle dans la littérature classique.

YADTK est une plateforme open-source pour développer des systèmes de dialogue oral. De part son caractère déclaratif et unifié, le modèle de représentation des connaissances permet un développement rapide et facilité.

Nous présentons TermLis un contexte d'information logique construit à partir de ressources terminologiques disponibles en xml (FranceTerme), pour une utilisation flexible avec un logiciel de contexte logique (CAMELIS). Une vue en contexte logique permet d'explorer des informations de manière flexible, sans rédaction de requête a priori, et d'obtenir aussi des indications sur la qualité des données. Un tel contexte peut être enrichi par d'autres informations (de natures diverses), mais aussi en le reliant à d'autres applications (par des actions associées selon des arguments fournis par le contexte). Nous montrons comment utiliser TermLis et nous illustrons, à travers cette réalisation concrète sur des données de FranceTerme, les avantages d'une telle approche pour des données terminologiques.

Ce travail concerne l'intégration à une plateforme de veille sur internet d'outils permettant l'analyse des opinions émises par les internautes à propos d'une entité, ainsi que la manière dont elles évoluent dans le temps. Les entités considérées peuvent être des personnes, des entreprises, des

marques, etc. Les outils implémentés sont le produit d'une collaboration impliquant plusieurs partenaires industriels et académiques dans le cadre du projet ANR ImagiWeb.

La création de ressources linguistiques de bonne qualité annotées en entités nommées est très coûteuse en temps et en main d'oeuvre. La plupart des corpus standards sont disponibles pour l'anglais mais pas pour les langues peu dotées, comme le vietnamien. Pour les langues asiatiques, cette tâche reste très difficile. Le présent article concerne la création automatique de corpus annotés en entités nommées pour le vietnamien-français, une paire de langues peu dotée. L'application d'une méthode basée sur la projection cross-lingue en utilisant des corpus parallèles. Les évaluations ont montré une bonne performance (F-score de 94.90%) lors de la reconnaissance des paires d'entités nommées dans les corpus parallèles et ainsi la construction d'un corpus bilingue annoté en entités nommées.

Nous présentons un outil en ligne de recherche de graphes dans des corpus annotés en syntaxe.

Le démonstrateur que nous décrivons ici est un prototype de système de dialogue dont l'objectif est de simuler un patient. Nous décrivons son fonctionnement général en insistant sur les aspects concernant la langue et surtout le rapport entre langue médicale de spécialité et langue générale.

Cette démonstration présente l'intégration du corpus arboré des Actes de TALN à la plateforme ScienQuest. Cette plateforme fut initialement créée pour l'étude du corpus de textes scientifiques Scientext. Cette intégration tient compte des méta-données propres au corpus TALN, et a été effectuée en s'efforçant de rapprocher les jeux d'étiquettes de ces deux corpus, et en convertissant pour le corpus TALN les requêtes prédéfinies conçues pour le corpus Scientext, de manière à permettre d'effectuer facilement des recherches similaires sur les deux corpus.

Cette démonstration présente une application mobile (pour tablette et smartphone) pour des personnes souffrant de troubles du langage et/ou de la parole permettant de générer des phrases à partir de la combinaison de pictogrammes puis de verbaliser le texte généré en Text-To-Speech (TTS). La principale critique adressée par les patients utilisant les solutions existantes est le temps de composition trop long d'une phrase. Cette limite ne permet pas ou très difficilement d'utiliser les solutions actuelles en condition dialogique. Pour pallier cela, nous avons développé un moteur de génération de texte avec prédiction sémantique ne proposant à l'utilisateur que les pictogrammes pertinents au regard de la saisie en cours (e.g. après le pictogramme [manger], l'application propose les pictogrammes [pomme] ou encore [viande] correspondant à des concepts comestibles). Nous avons ainsi multiplié de 5 à 10 la vitesse de composition d'une phrase par rapport aux solutions existantes.

Le projet européen TIER (Integrated strategy for CBRN ? Chemical, Biological, Radiological and Nuclear ? Threat Identification and Emergency Response) vise à intégrer une stratégie complète et intégrée pour la réponse d'urgence dans un contexte de dangers biologiques, chimiques, radiologiques, nucléaires, ou liés aux explosifs, basée sur l'identification des menaces et d'évaluation des risques. Dans cet article, nous nous focalisons sur les risques biologiques. Nous présentons notre système expert fondé sur une analyse sémantique, permettant l'extraction de données structurées à partir de données non structurées dans le but de raisonner.

Dans cette démonstration, nous présentons l'annotateur multi-niveaux DisMo, un outil conçu pour faire face aux spécificités des corpus oraux. Il fournit une annotation morpho-syntaxique, une lemmatisation, une détection des unités poly-lexicales, une détection des phénomènes de disfluence et des marqueurs de discours.

Dans cette présentation, je défendrai l'idée selon laquelle des ressources terminologiques décrivant

les propriétés lexico-sémantiques des termes constituent un complément nécessaire, voire indispensable, à d'autres types de ressources, À partir d'exemples anglais et français empruntés au domaine de l'environnement, je montrerai, d'une part, que les ressources lexicales générales (y compris celles qui ont une large couverture) n'offrent pas un portait complet du sens des termes ou de la structure lexicale observée du point de vue d'un domaine de spécialité. Je montrerai, d'autre part, que les ressources terminologiques (thésaurus, ontologies, banques de terminologie) souvent d'obédience conceptuelle, se concentrent sur le lien entre les termes et les connaissances dénotées par eux et s'attardent peu sur leur fonctionnement linguistique. Je présenterai un type de ressource décrivant les propriétés lexico-sémantiques des termes d'un domaine (structure actantielle, liens lexicaux, annotations contextuelles, etc.) et des éléments méthodologiques présidant à son élaboration.

De nombreuses méthodes ont été proposées pour accélérer la prédiction d'objets structurés (tels que les arbres ou les séquences), ou pour permettre la prise en compte de dépendances plus riches afin d'améliorer les performances de la prédiction. Ces méthodes reposent généralement sur des techniques d'inférence approchée et ne bénéficient d'aucune garantie théorique aussi bien du point de vue de la qualité de la solution trouvée que du point de vue de leur critère d'apprentissage. Dans ce travail, nous étudions une nouvelle formulation de l'apprentissage structuré qui consiste à voir celui-ci comme un processus incrémental au cours duquel la sortie est construite de façon progressive. Ce cadre permet de formaliser plusieurs approches de prédiction structurée existantes. Grâce au lien que nous faisons entre apprentissage structuré et apprentissage par renforcement, nous sommes en mesure de proposer une méthode théoriquement bien justifiée pour apprendre des méthodes d'inférence approchée. Les expériences que nous réalisons sur quatre tâches de TAL valident l'approche proposée.

Beaucoup de problèmes de TAL sont désormais modélisés comme des tâches d'apprentissage

supervisé. De ce fait, le coût des annotations des exemples par l'expert représente un problème important. L'apprentissage actif (active learning) apporte un cadre à ce problème, permettant de contrôler le coût d'annotation tout en maximisant, on l'espère, la performance de la tâche visée, mais repose sur le choix difficile des exemples à soumettre à l'expert. Dans cet article, nous examinons et proposons des stratégies de sélection des exemples pour le cas spécifique des champs aléatoires conditionnels (Conditional Random Fields, CRF), outil largement utilisé en TAL. Nous proposons d'une part une méthode simple corrigeant un biais de certaines méthodes de l'état de l'art. D'autre part, nous détaillons une méthode originale de sélection s'appuyant sur un critère de respect des proportions dans les jeux de données manipulés. Le bien- fondé de ces propositions est vérifié au travers de plusieurs tâches et jeux de données, incluant reconnaissance d'entités nommées, chunking, phonétisation, désambiguïsation de sens.

Dans cet article, nous présentons les méthodes que nous avons développées pour analyser des comptes- rendus hospitaliers rédigés en anglais. L'objectif de cette étude consiste à identifier les facteurs de risque de décès pour des patients diabétiques et à positionner les événements médicaux décrits par rapport à la date de création de chaque document. Notre approche repose sur (i) HeidelTime pour identifier les expressions temporelles, (ii) des CRF complétés par des règles de post-traitement pour identifier les traitements, les maladies et facteurs de risque, et (iii) des règles pour positionner temporellement chaque événement médical. Sur un corpus de 514 documents, nous obtenons une F-mesure globale de 0,8451. Nous observons que l'identification des informations directement mentionnées dans les documents se révèle plus performante que l'inférence d'informations à partir de résultats de laboratoire.

Quand on dispose de connaissances a priori sur les sorties possibles d'un problème d'étiquetage, il semble souhaitable d'inclure cette information lors de l'apprentissage pour simplifier la tâche de modélisation et accélérer les traitements. Pourtant, même lorsque ces contraintes sont correctes et

utiles au décodage, leur utilisation lors de l'apprentissage peut dégrader sévèrement les performances. Dans cet article, nous étudions ce paradoxe et montrons que le manque de contraste induit par les connaissances entraîne une forme de sous-apprentissage qu'il est cependant possible de limiter.

Les références à des phénomènes du monde réel et à leur caractérisation temporelle se retrouvent dans beaucoup de types de discours en langue naturelle. Ainsi, l'analyse temporelle apparaît comme un élément important en traitement automatique de la langue. Cet article présente une analyse de textes en domaine de spécialité du point de vue temporel. En s'appuyant sur un corpus de documents issus de plusieurs dossiers électroniques patient désidentifiés, nous décrivons la construction d'une ressource annotée en expressions temporelles selon la norme TimeML. Par suite, nous utilisons cette ressource pour évaluer plusieurs méthodes d'extraction automatique d'expressions temporelles adaptées au domaine médical. Notre meilleur système statistique offre une performance de 0,91 de F-mesure, surpassant pour l'identification le système état de l'art HeidelTime. La comparaison de notre corpus de travail avec le corpus journalistique FR-Timebank permet également de caractériser les différences d'utilisation des expressions temporelles dans deux domaines de spécialité.

La majorité des méthodes état de l'art en compréhension automatique de la parole ont en commun de devoir être apprises sur une grande quantité de données annotées. Cette dépendance aux données constitue un réel obstacle lors du développement d'un système pour une nouvelle tâche/langue. Aussi, dans cette étude, nous présentons une méthode visant à limiter ce besoin par un mécanisme d'apprentissage sans données de référence (zero-shot learning). Cette méthode combine une description ontologique minimale de la tâche visée avec l'utilisation d'un espace sémantique continu appris par des approches à base de réseaux de neurones à partir de données génériques non-annotées. Nous montrons que le modèle simple et peu coûteux obtenu peut

atteindre, dès le démarrage, des performances comparables à celles des systèmes état de l'art reposant sur des règles expertes ou sur des approches probabilistes sur des tâches de compréhension de la parole de référence (tests des Dialog State Tracking Challenges, DSTC2 et DSTC3). Nous proposons ensuite une stratégie d'adaptation en ligne permettant d'améliorer encore les performances de notre approche à l'aide d'une supervision faible et ajustable par l'utilisateur.

Cet article présente un modèle génératif pour l'induction non supervisée d'événements. Les précédentes méthodes de la littérature utilisent uniquement les têtes des syntagmes pour représenter les entités. Pourtant, le groupe complet (par exemple, "un homme armé") apporte une information plus discriminante (que "homme"). Notre modèle tient compte de cette information et la représente dans la distribution des schémas d'événements. Nous montrons que ces relations jouent un rôle important dans l'estimation des paramètres, et qu'elles conduisent à des distributions plus cohérentes et plus discriminantes. Les résultats expérimentaux sur le corpus de MUC-4 confirment ces progrès.

Nous présentons une méthode pour créer rapidement un système de désambiguïsation lexicale (DL) pour une langue L peu dotée pourvu que l'on dispose d'un système de traduction automatique statistique (TAS) d'une langue riche en corpus annotés en sens (ici l'anglais) vers L. Il est, en effet, plus facile de disposer des ressources nécessaires à la création d'un système de TAS que des ressources dédiées nécessaires à la création d'un système de DL pour la langue L. Notre méthode consiste à traduire automatiquement un corpus annoté en sens vers la langue L, puis de créer le système de désambiguïsation pour L par des méthodes supervisées classiques. Nous montrons la faisabilité de la méthode et sa généricité en traduisant le SemCor, un corpus en anglais annoté grâce au Princet-on Word-Net, de l'anglais vers le bangla et de l'anglais vers le français. Nous montrons la validité de l'approche en évaluant les résultats sur la tâche de désambiguïsation lexicale multi-lingue de Semeval 2013.

La détection d'opinion ciblée a pour but d'attribuer une opinion à une caractéristique particulière d'un produit donné. La plupart des méthodes existantes envisagent pour cela une approche non supervisée. Or, les utilisateurs ont souvent une idée a priori des caractéristiques sur lesquelles ils veulent découvrir l'opinion des gens. Nous proposons dans cet article une méthode pour une extraction d'opinion ciblée, qui exploite cette information minimale sur les caractéristiques d'intérêt. Ce modèle s'appuie sur une segmentation automatique des textes, un enrichissement des données disponibles par similarité sémantique, et une annotation de l'opinion par classification supervisée. Nous montrons l'intérêt de l'approche sur un cas d'étude dans le domaine des jeux vidéos.

Cet article entend dresser, dans un premier temps, un panorama critique des relations entre TAL et linguistique. Puis, il esquisse une discussion sur l'apport possible d'une sémantique de corpus dans un contexte applicatif en s'appuyant sur plusieurs expériences en fouille de textes subjectifs (analyse de sentiments et fouille d'opinions). Ces expériences se démarquent des approches traditionnelles fondées sur la recherche de marqueurs axiologiques explicites par l'utilisation de critères relevant des représentations des acteurs (composante dialogique) et des structures argumentatives et narratives des textes (composante dialectique). Nous souhaitons de cette façon mettre en lumière le bénéfice d'un dialogue méthodologique entre une théorie (la sémantique textuelle), des méthodes de linguistique de corpus orientées vers l'analyse du sens (la textométrie) et les usages actuels du TAL en termes d'algorithmiques (apprentissage automatique) mais aussi de méthodologie d'évaluation des résultats.

L'homogénéité sémantique stipule que des termes sont sémantiquement proches mais non similaires. Cette notion est au coeur de travaux relatifs à la génération automatique de questionnaires à choix multiples, et particulièrement à la sélection automatique de distracteurs. Dans cet article, nous présentons une méthode d'estimation de l'homogénéité sémantique dans un

cadre de validation automatique de distracteurs. Cette méthode est fondée sur une combinaison de plusieurs critères de voisinage et de similarité sémantique entre termes, par apprentissage automatique. Nous montrerons que notre méthode permet d'obtenir une meilleure estimation de l'homogénéité sémantique que les méthodes proposées dans l'état de l'art.

Cet article présente une expérimentation visant à construire une ressource sémantique pour le français contemporain à partir d'un corpus d'environ un million de définitions tirées de deux ressources lexicographiques (Trésor de la Langue Française, Wiktionary) et d'une ressource encyclopédique (Wikipedia). L'objectif est d'extraire automatiquement dans les définitions différentes relations sémantiques : hyperonymie, synonymie, méronymie, autres relations sémantiques. La méthode suivie combine la précision des patrons lexico-syntaxiques et le rappel des méthodes statistiques, ainsi qu'un traitement inédit de canonisation et de décomposition des énoncés. Après avoir présenté les différentes approches et réalisations existantes, nous détaillons l'architecture du système et présentons les résultats : environ 900 000 relations d'hyperonymie et près de 100 000 relations de synonymie, avec un taux de précision supérieur à 90% sur un échantillon aléatoire de 500 relations. Plus de 2 millions de prédications définitoires ont également été extraites.

La plupart des méthodes d'amélioration des thésaurus distributionnels se focalisent sur les moyens ? représentations ou mesures de similarité ? de mieux détecter la similarité sémantique entre les mots. Dans cet article, nous proposons un point de vue inverse : nous cherchons à détecter les voisins sémantiques associés à une entrée les moins susceptibles d'être liés sémantiquement à elle et nous utilisons cette information pour réordonner ces voisins. Pour détecter les faux voisins sémantiques d'une entrée, nous adoptons une approche s'inspirant de la désambiguïsation sémantique en construisant un classifieur permettant de différencier en contexte cette entrée des autres mots. Ce classifieur est ensuite appliqué à un échantillon des occurrences des voisins de l'entrée pour repérer ceux les plus éloignés de l'entrée. Nous évaluons cette méthode pour des

thésaurus construits à partir de co-occurrents syntaxiques et nous montrons l'intérêt de la combiner avec les méthodes décrites dans (Ferret, 2013b) selon une stratégie de type vote.

Nous présentons une méthode pour articuler grammaire de phrase et grammaire de discours qui évite de recourir à une étape de traitement intermédiaire. Cette méthode est suffisamment générale pour construire des structures discursives qui ne soient pas des arbres mais des graphes orientés acycliques (DAG). Notre analyse s'appuie sur une approche de l'analyse discursive, Discourse Synchronous TAG (D-STAG), qui utilise les Grammaires d'Arbres Adjoint (TAG). Nous utilisons pour ce faire un encodage des TAG dans les Grammaires Catégorielles Abstraites (ACG). Cet encodage permet d'une part d'utiliser l'ordre supérieur pour l'interprétation sémantique afin de construire des structures qui soient des DAG et non des arbres, et d'autre part d'utiliser les propriétés de composition d'ACG pour réaliser naturellement l'interface entre grammaire phrastique et grammaire discursive. Tous les exemples proposés pour illustrer la méthode ont été implantés et peuvent être testés avec le logiciel approprié.

De nombreux problèmes en traitement automatique des langues requièrent de déterminer si deux phrases sont des réécritures l'une de l'autre. Une solution efficace consiste à apprendre les réécritures en se fondant sur des méthodes à noyau qui mesurent la similarité entre deux réécritures de paires de phrases. Toutefois, ces méthodes ne permettent généralement pas de prendre en compte des variations sémantiques entre mots, qui permettraient de capturer un plus grand nombre de règles de réécriture. Dans cet article, nous proposons la définition et l'implémentation d'une nouvelle classe de fonction noyau, fondée sur la réécriture de phrases enrichie par un typage pour combler ce manque. Nous l'évaluons sur deux tâches, la reconnaissance de paraphrases et d'implications textuelles.

Nous sommes tous concernés par notre état de santé et restons sensibles aux informations de

santé disponibles dans la société moderne à travers par exemple les résultats des recherches scientifiques, les médias sociaux de santé, les documents cliniques, les émissions de télé et de radio ou les nouvelles. Cependant, il est commun de rencontrer dans le domaine médical des termes très spécifiques (e.g., blépharospasme, alexitymie, appendicectomie), qui restent difficiles à comprendre par les non spécialistes. Nous proposons une méthode automatique qui vise l'acquisition de paraphrases pour les termes médicaux, qui soient plus faciles à comprendre que les termes originaux. La méthode est basée sur l'analyse morphologique des termes, l'analyse syntaxique et la fouille de textes non spécialisés. L'analyse et l'évaluation des résultats indiquent que de telles paraphrases peuvent être trouvées dans les documents non spécialisés et présentent une compréhension plus facile. En fonction des paramètres de la méthode, la précision varie entre 86 et 55 %. Ce type de ressources est utile pour plusieurs applications de TAL (e.g., recherche d'information grand public, lisibilité et simplification de textes, systèmes de question-réponses).

L'article présente des résultats d'expériences d'apprentissage automatique pour l'étiquetage morpho-syntaxique et l'analyse syntaxique en dépendance de l'ancien français. Ces expériences ont pour objectif de servir une exploration de corpus pour laquelle le corpus arboré SRCMF sert de données de référence. La nature peu standardisée de la langue qui y est utilisée implique des données d'entraînement hétérogènes et quantitativement limitées. Nous explorons donc diverses stratégies, fondées sur différents critères (variabilité du lexique, forme Vers/Prose des textes, dates des textes), pour constituer des corpus d'entrainement menant aux meilleurs résultats possibles.

Cet article s'attaque à la tâche d'Attribution d'Auteur en contexte multi-lingue. Nous proposons une alternative aux méthodes supervisées fondées sur les n-grammes de caractères de longueurs variables : les répétitions maximales. Pour un texte donné, la liste de ses n-grammes de caractères contient des informations redondantes. A contrario, les répétitions maximales représentent l'ensemble des répétitions de ce texte de manière condensée. Nos expériences montrent que la

redondance des n-grammes contribue à l'efficacité des techniques d'Attribution d'Auteur exploitant des sous-chaînes de caractères. Ce constat posé, nous proposons une fonction de pondération sur les traits donnés en entrée aux classifieurs, en introduisant les répétitions maximales du nème ordre (c'est-à-dire des répétitions maximales détectées dans un ensemble de répétitions maximales). Les résultats expérimentaux montrent de meilleures performances avec des répétitions maximales, avec moins de données que pour les approches fondées sur les n-grammes.

Cet article présente une méthode pour mesurer la similarité sémantique entre phrases qui utilise Wikipédia comme unique ressource linguistique et qui est, de ce fait, utilisable pour un grand nombre de langues. Basée sur une représentation vectorielle, elle utilise une indexation aléatoire pour réduire la dimension des espaces manipulés. En outre, elle inclut une technique de calcul des vecteurs de termes qui corrige les défauts engendrés par l'utilisation d'un corpus aussi général que Wikipédia. Le système a été évalué sur les données de SemEval 2014 en anglais avec des résultats très encourageants, au-dessus du niveau moyen des systèmes en compétition. Il a également été testé sur un ensemble de paires de phrases en français, à partir de ressources que nous avons construites et qui seront mises à la libre disposition de la communauté scientifique.

La typologie des langues repose sur l'étude de la réalisation de propriétés ou phénomènes linguistiques dans plusieurs langues ou familles de langues. Nous abordons dans cet article la question de la typologie syntaxique et proposons une méthode permettant d'extraire automatiquement ces propriétés à partir de treebanks, puis de les analyser en vue de dresser une telle typologie. Nous décrivons cette méthode ainsi que les outils développés pour la mettre en oeuvre. Celle-ci a été appliquée à l'analyse de 10 langues décrites dans le Universal Dependencies Treebank. Nous validons ces résultats en montrant comment une technique de classification permet, sur la base des informations extraites, de reconstituer des familles de langues.

Les mesures de confiance au niveau mot (Word Confidence Estimation - WCE) pour la traduction auto- matique (TA) ou pour la reconnaissance automatique de la parole (RAP) attribuent un score de confiance à chaque mot dans une hypothèse de transcription ou de traduction. Dans le passé, l'estimation de ces mesures a le plus souvent été traitée séparément dans des contextes RAP ou TA. Nous proposons ici une estimation conjointe de la confiance associée à un mot dans une hypothèse de traduction automatique de la parole (TAP). Cette estimation fait appel à des paramètres issus aussi bien des systèmes de transcription de la parole (RAP) que des systèmes de traduction automatique (TA). En plus de la construction de ces estimateurs de confiance robustes pour la TAP, nous utilisons les informations de confiance pour re-décoder nos graphes d'hypothèses de traduction. Les expérimentations réalisées montrent que l'utilisation de ces mesures de confiance au cours d'une seconde passe de décodage permettent d'obtenir une amélioration significative des performances de traduction (évaluées avec la métrique BLEU - gains de deux points par rapport à notre système de traduc- tion de parole de référence). Ces expériences sont faites pour une tâche de TAP (français-anglais) pour laquelle un corpus a été spécialement conçu (ce corpus, mis à la disposition de la communauté TALN, est aussi décrit en détail dans l'article).

Dans cet article, nous proposons une méthode originale destinée à effectuer l'alignement d'un corpus multi-parallèle, i.e. comportant plus de deux langues, en prenant en compte toutes les langues simultanément (et non en composant une série de bi-alignements indépendants). Pour ce faire, nous nous appuyons sur les réseaux de correspondances lexicales constitués par les transfuges (chaînes identiques) et cognats (mots apparentés), et nous montrons comment divers tuilages des couples de langues permettent d'exploiter au mieux les ressemblances superficielles liées aux relations génétiques interlinguistiques. Nous évaluons notre méthode par rapport à une méthode de bi-alignement classique, et montrons en quoi le multi-alignement permet d'obtenir des résultats à la fois plus précis et plus robustes.

Alors que les réseaux neuronaux occupent une place de plus en plus importante dans le traitement automatique des langues, les méthodes d'apprentissage actuelles utilisent pour la plupart des critères qui sont décorrélés de l'application. Cet article propose un nouveau cadre d'apprentissage discriminant pour l'estimation des modèles continus de traduction. Ce cadre s'appuie sur la définition d'un critère d'optimisation permettant de prendre en compte d'une part la métrique utilisée pour l'évaluation de la traduction et d'autre part l'intégration de ces modèles au sein des systèmes de traduction automatique. De plus, cette méthode d'apprentissage est comparée aux critères existants d'estimation que sont le maximum de vraisemblance et l'estimation contrastive bruitée. Les expériences menées sur la tâches de traduction des sémin-aires TED Talks de l'anglais vers le français montrent la pertinence d'un cadre discriminant d'apprentissage, dont les performances restent toutefois très dépendantes du choix d'une stratégie d'initialisation idoine. Nous montrons qu'avec une initialisation judicieuse des gains significatifs en termes de scores BLEU peuvent être obtenus.

Bien que les interjections soient un phénomène linguistique connu, elles ont été peu étudiées et cela continue d'être le cas pour les travaux sur les microblogs. Des travaux en analyse de sentiments ont montré l'intérêt des émoticônes et récemment des mots-dièses, qui s'avèrent être très utiles pour la classification en polarité. Mais malgré leur statut grammatical et leur richesse sémantique, les interjections sont restées marginalisées par les systèmes d'analyse de sentiments. Nous montrons dans cet article l'apport majeur des interjections pour la détection des émotions. Nous détaillons la production automatique, basée sur les interjections, d'un corpus étiqueté avec les émotions. Nous expliquons ensuite comment nous avons utilisé ce corpus pour en déduire, automatiquement, un lexique affectif pour le français. Ce lexique a été évalué sur une tâche de détection des émotions, qui a montré un gain en mesure F1 allant, selon les émotions, de +0,04 à +0,21.

L'article traite de l'analyse syntaxique lexicalisée pour les grammaires de constituants. On se place dans le cadre de l'analyse par transitions. Les modèles statistiques généralement utilisés pour cette tâche s'appuient sur une représentation non structurée du lexique. Les mots du vocabulaire sont représentés par des symboles discrets sans liens entre eux. À la place, nous proposons d'utiliser des représentations denses du type plongements (embeddings) qui permettent de modéliser la similarité entre symboles, c'est-à-dire entre mots, entre parties du discours et entre catégories syntagmatiques. Nous proposons d'adapter le modèle statistique sous-jacent à ces nouvelles représentations. L'article propose une étude de 3 architectures neuronales de complexité croissante et montre que l'utilisation d'une couche cachée non-linéaire permet de tirer parti des informations données par les plongements.

Notre travail porte sur la détection automatique des segments en relation de reformulation paraphrastique dans les corpus oraux. L'approche proposée est une approche syntagmatique qui tient compte des marqueurs de reformulation paraphrastique et des spécificités de l'oral. Les données de référence sont consensuelles. Une méthode automatique fondée sur l'apprentissage avec les CRF est proposée afin de détecter les segments paraphrasés. Différents descripteurs sont exploités dans une fenêtre de taille variable. Les tests effectués montrent que les segments en relation de paraphrase sont assez difficiles à détecter, surtout avec leurs frontières correctes. Les meilleures moyennes atteignent 0,65 de F-mesure, 0,75 de précision et 0,63 de rappel. Nous avons plusieurs perspectives à ce travail pour améliorer la détection des segments en relation de paraphrase et pour étudier les données depuis d'autres points de vue.

Cet article présente une approche associant réseaux lexico-sémantiques et représentations distribuées

de mots appliquée à l'évaluation de la traduction automatique. Cette étude est faite à travers

l'enrichissement d'une métrique bien connue pour évaluer la traduction automatique (TA) : METEOR.

METEOR permet un appariement approché (similarité morphologique ou synonymie) entre une sortie de système automatique et une traduction de référence. Nos expérimentations s'appuient sur la tâche Metrics de la campagne d'évaluation WMT 2014 et montrent que les représentations distribuées restent moins performantes que les ressources lexico-sémantiques pour l'évaluation en TA
mais peuvent néammoins apporter un complément d'information intéressant à ces dernières.

Dans cet article nous évaluons l'impact de la prise en compte du contexte dans la traduction de dialogues. Nous introduisons pour cela un nouveau corpus parallèle, issu des sous-titres de séries télévisées, comportant de nombreuses informations contextuelles. Nous montrons comment la prise en compte du genre du locuteur permet d'améliorer significativement la qualité de la traduction automatique en termes de score BLEU et METEOR . Une analyse manuelle montre toutefois que ces
gains ne sont pas nécessairement liés aux conséquences morphologiques du genre du locuteur, mais à
des différences linguistiques plus générales.

Dans certains textes bruts, les marques de fin de ligne peuvent marquer ou pas la frontière d'une unité
textuelle (typiquement un paragraphe). Ce problème risque d'influencer les traitements subséquents,
mais est rarement traité dans la littérature. Nous proposons une méthode entièrement non-supervisée
pour déterminer si une fin de ligne doit être vue comme un simple espace ou comme une véritable

frontière d'unité textuelle, et la testons sur un corpus de comptes rendus médicaux. Cette méthode obtient une F-mesure de 0,926 sur un échantillon de 24 textes contenant des lignes repliées. Appliquée
sur un échantillon plus grand de textes contenant ou pas des lignes repliées, notre méthode la plus prudente obtient une F-mesure de 0,898, valeur élevée pour une méthode entièrement non-supervisée.

Nous présentons, dans cet article, une adaptation d'un processus d'extraction de termes pour l'arabe
standard moderne. L'adaptation a d'abord consisté à décrire le processus d'extraction des termes de
manière similaire à celui défini pour l'anglais et le français en prenant en compte certaines particularités morpho-syntaxiques de la langue arabe. Puis, nous avons considéré le phénomène de l'agglutination de la langue arabe. L'évaluation a été réalisée sur un corpus de textes médicaux. Les résultats
montrent que parmi 400 termes candidats maximaux analysés, 288 sont jugés corrects par rapport au domaine (72,1%). Les erreurs d'extraction sont dues à l'étiquetage morpho-syntaxique et à la non-voyellation des textes mais aussi à complexité de la prise en compte de l'agglutination.

Dans le contexte des masses de données aujourd'hui disponibles, de nombreux travaux liés à l'analyse
de l'information spatiale s'appuient sur l'exploitation des données textuelles. La communication médiée (SMS, tweets, etc.) véhiculant des informations spatiales prend une place prépondérante. L'objectif du travail présenté dans cet article consiste à extraire ces informations spatiales à partir d'un corpus authentique de SMS en français. Nous proposons un processus dans lequel, dans un premier temps, nous extrayons de nouvelles entités spatiales (par exemple, motpellier, montpeul à

associer au toponyme Montpellier). Dans un second temps, nous identifions de nouvelles relations spatiales qui précèdent les entités spatiales (par exemple, sur, par, pres, etc.). La tâche est difficile et
complexe en raison de la spécificité du langage SMS qui repose sur une écriture peu standardisée (apparition de nombreux lexiques, utilisation massive d'abréviations, variation par rapport à l'écrit classique, etc.). Les expérimentations qui ont été réalisées à partir du corpus 88milSMS mettent en relief la robustesse de notre système pour identifier de nouvelles entités et relations spatiales.

Pour un certain nombre de tâches ou d'applications du TALN, il est nécessaire de déterminer la proximité sémantique entre des sens, des mots ou des segments textuels. Dans cet article, nous nous
intéressons à une mesure basée sur des savoirs, la mesure de Lesk. La proximité sémantique de deux
définitions est évaluée en comptant le nombre de mots communs dans les définitions correspondantes
dans un dictionnaire. Dans cet article, nous étudions plus particulièrement l'extension de définitions grâce à des corpus annotés en sens. Il s'agit de prendre en compte les mots qui sont utilisés dans le
voisinage d'un certain sens et d'étendre lexicalement la définition correspondante. Nous montrons une amélioration certaine des performances obtenues en désambiguïsation lexicale qui dépassent l'état de l'art.

Cet article présente une méthode de construction d'une ressource lexicale de sentiments/émotions. Son originalité est d'associer le crowdsourcing via un GWAP (Game With A Purpose) à un algorithme de propagation, les deux ayant pour support et source de données le réseau lexical JeuxDeMots. Nous décrivons le jeu permettant de collecter des informations de sentiments, ainsi

que les principes et hypothèses qui sous-tendent le fonctionnement de l'algorithme qui les propage au sein du réseau. Enfin, nous donnons les résultats quantitatifs et expliquons les méthodes d'évaluation qualitative des données obtenues, à la fois par le jeu et par la propagation par l'algorithme. Ces méthodes incluent une comparaison avec Emolex, une autre ressource de sentiments/émotions.

Dans cet article, nous évaluons, à travers son intérêt pour le résumé automatique et la détection d'ancres dans des vidéos, le potentiel d'une nouvelle structure thématique extraite de données textuelles, composée d'une hiérarchie de fragments thématiquement focalisés. Cette structure est produite par un algorithme exploitant les distributions temporelles d'apparition des mots dans les textes en se fondant sur une analyse de salves lexicales. La hiérarchie obtenue a pour objet de filtrer
le contenu non crucial et de ne conserver que l'information saillante des textes, à différents niveaux de détail. Nous montrons qu'elle permet d'améliorer la production de résumés ou au moins de maintenir les résultats de l'état de l'art, tandis que pour la détection d'ancres, elle nous conduit à la meilleure précision dans le contexte de la tâche Search and Anchoring in Video Archives à MediaEval.
Les expériences sont réalisées sur du texte écrit et sur un corpus de transcriptions automatiques d'émissions de télévision.

Les notions de domaines techniques, comme les notions médicales, présentent souvent des difficultés
de compréhension par les non experts. Un vocabulaire qui associe les termes techniques aux expressions grand public peut aider à rendre les textes techniques mieux compréhensibles. L'objectif de notre
travail est de construire un tel vocabulaire. Nous proposons d'exploiter la notion de reformulation

grâce à trois méthodes : extraction d'abréviations, exploitation de marqueurs de reformulation et de parenthèses. Les segments associés grâce à ces méthodes sont alignés avec les terminologies médicales. Nos résultats permettent de couvrir un grand nombre de termes médicaux et montrent une précision d'extraction entre 0,68 et 0,98. Au total, plusieurs dizaines de milliers de paires sont proposés. Ces résultats sont analysés et comparés avec les travaux existants.

L'identification des entités nommées dans un texte est une tâche essentielle des outils d'extraction d'information dans de nombreuses applications. Cette identification passe par la reconnaissance d'une
mention d'entité dans le texte, ce qui a été très largement étudié, et par l'association des entités reconnues à des entités connues, présentes dans une base de connaissances. Cette association repose
souvent sur une mesure de similarité entre le contexte textuel de la mention de l'entité et un contexte
textuel de description des entités de la base de connaissances. Or, ce contexte de description n'est en général pas présent pour toutes les entités. Nous proposons d'exploiter les relations de la base de connaissances pour ajouter un indice de désambiguïsation pour ces entités. Nous évaluons notre
travail sur des corpus d'évaluation standards en anglais issus de la tâche de désambiguïsation d'entités
de la campagne TAC-KBP.

Cet article présente un modèle bayésien non-paramétrique pour la segmentation morphologique non supervisée. Ce modèle semi-markovien s'appuie sur des classes latentes de morphèmes afin de
modéliser les caractéristiques morphotactiques du lexique, et son caractère non-paramétrique lui

permet de s'adapter aux données sans avoir à spécifier à l'avance l'inventaire des morphèmes ainsi que leurs classes. Un processus de Pitman-Yor est utilisé comme a priori sur les paramètres afin d'éviter une convergence vers des solutions dégénérées et inadaptées au traitemement automatique
des langues. Les résultats expérimentaux montrent la pertinence des segmentations obtenues pour le turc et l'anglais. Une étude qualitative montre également que le modèle infère une morphotactique linguistiquement pertinente, sans le recours à des connaissances expertes quant à la structure
morphologique des formes de mots.

La traduction automatique statistique bien que performante est aujourd'hui limitée parce qu'elle nécessite de gros volumes de corpus parallèles qui n'existent pas pour tous les couples de langues et
toutes les spécialités et que leur production est lente et coûteuse. Nous présentons, dans cet article, un prototype d'un moteur de traduction à base d'exemples utilisant la recherche d'information interlingue et ne nécessitant qu'un corpus de textes en langue cible. Plus particulièrement, nous proposons d'étudier l'impact d'un lexique bilingue de spécialité sur la performance de ce prototype. Nous évaluons ce prototype de traduction et comparons ses résultats à ceux du système de traduction
statistique Moses en utilisant les corpus parallèles anglais-français Europarl (European Parliament Proceedings) et Emea (European Medicines Agency Documents). Les résultats obtenus montrent que le score BLEU du prototype du moteur de traduction à base d'exemples est proche de celui du système Moses sur des documents issus du corpus Europarl et meilleur sur des documents extraits du corpus Emea.

Dans cet article nous étudions plusieurs types de réseaux neuronaux récurrents (RNN) pour

l'étiquetage de séquences. Nous proposons deux nouvelles variantes de RNN et nous les comparons
aux variantes plus classiques de type Jordan et Elman. Nous expliquons en détails quels sont les avantages de nos nouvelles variantes par rapport aux autres RNN. Nous évaluons tous les modèles,
les nouvelles variantes ainsi que les RNN existants, sur deux tâches de compréhension de la parole :
ATIS et MEDIA. Les résultats montrent que nos nouvelles variantes de RNN sont plus efficaces que les autres.

Cette étude examine l'utilisation de méthodes d'apprentissage incrémental supervisé afin de prédire la compétence lexicale d'apprenants de français langue étrangère (FLE). Les apprenants ciblés sont des néerlandophones ayant un niveau A2/B1 selon le Cadre européen commun de référence pour les
langues (CECR). À l'instar des travaux récents portant sur la prédiction de la maîtrise lexicale à l'aide d'indices de complexité, nous élaborons deux types de modèles qui s'adaptent en fonction d'un
retour d'expérience, révélant les connaissances de l'apprenant. En particulier, nous définissons (i) un
modèle qui prédit la compétence lexicale de tous les apprenants du même niveau de maîtrise et (ii) un modèle qui prédit la compétence lexicale d'un apprenant individuel. Les modèles obtenus sont ensuite évalués par rapport à un modèle de référence déterminant la compétence lexicale à partir d'un
lexique spécialisé pour le FLE et s'avèrent gagner significativement en exactitude (9%-17%).

Dans cet article, nous nous intéressons à l'extraction d'entités médicales de type symptôme dans

les textes biomédicaux. Cette tâche est peu explorée dans la littérature et il n'existe pas à notre connaissance de corpus annoté pour entraîner un modèle d'apprentissage. Nous proposons deux approches faiblement supervisées pour extraire ces entités. Une première est fondée sur la fouille de motifs et introduit une nouvelle contrainte de similarité sémantique. La seconde formule la tache comme une tache d'étiquetage de séquences en utilisant les CRF (champs conditionnels aléatoires). Nous décrivons les expérimentations menées qui montrent que les deux approches sont complémentaires en termes d'évaluation quantitative (rappel et précision). Nous montrons en outre que leur combinaison améliore sensiblement les résultats.

Ce papier décrit une approche pour créer des résumés de conversations parlées par remplissage de
patrons. Les patrons sont générés automatiquement à partir de fragments généralisés depuis un corpus
de résumés d'apprentissage. Les informations nécessaires pour remplir les patrons sont détectées dans les transcriptions des conversations et utilisées pour sélectionner les fragments candidats. L'approche obtient un score ROUGE-2 de 0.116 sur le corpus RATP-DECODA. Les résultats obtenus
montrent que cette approche abstractive est plus performante que les approches extractives utilisées
habituellement dans le domaine du résumé automatique.

La reformulation participe à la structuration du discours, notamment dans le cas des dialogues, et contribue également à la dynamique du discours. Reformuler est un acte significatif qui poursuit des objectifs précis. L'objectif de notre travail est de prédire automatiquement la raison pour laquelle un locuteur effectue une reformulation. Nous utilisons une classification de onze fonctions pragmatiques

inspirées des travaux existants et des données analysées. Les données de référence sont issues d'annotations manuelles et consensuelles des reformulations spontanées formées autour de trois marqueurs (c'est-à-dire, je veux dire, disons). Les données proviennent d'un corpus oral et d'un corpus de discussions sur les forums de santé. Nous exploitons des algorithmes de catégorisation supervisée et un ensemble de plusieurs descripteurs (syntaxiques, formels, sémantiques et discursifs)
pour prédire les catégories de reformulation. La distribution des énoncés et phrases selon les catégories
n'est pas homogène. Les expériences sont positionnées à deux niveaux : générique et spécifique. Nos résultats indiquent qu'il est plus facile de prédire les types de fonctions au niveau générique (la moyenne des F-mesures est autour de 0,80), qu'au niveau des catégories individuelles (la moyenne des F-mesures est autour de 0,40). L'influence de différents paramètres est étudiée.

L'analyse des conversations écrites porteuses de demandes d'assistance est un enjeu important pour le développement de nouvelles technologies liées au support client. Dans cet article, nous nous intéressons à l'analyse d'un même type d'échange sur un canal différent : les conversations se déroulant sur les plate-formes d'entraide entre utilisateurs. Nous comparons des approches de classification supervisées sur trois modalités des CMR 1 différentes à même thématique : des courriels,
forums et chats issus de la communauté Ubuntu. Le système emploie une taxonomie fine basée sur le
schéma DIT++. D'autres expériences sont détaillées, et nous rapportons les résultats obtenus avec différentes approches et différents traits sur les différentes parties de notre corpus multimodal.

Dans cet article, nous nous intéressons à l'indexation de documents de domaines de spécialité par l'intermédiaire de leurs termes-clés. Plus particulièrement, nous nous intéressons à l'indexation

telle qu'elle est réalisée par les documentalistes de bibliothèques numériques. Après analyse de la méthodologie de ces indexeurs professionnels, nous proposons une méthode à base de graphe combinant les informations présentes dans le document et la connaissance du domaine pour réaliser une indexation (hybride) libre et contrôlée. Notre méthode permet de proposer des termes-clés ne se trouvant pas nécessairement dans le document. Nos expériences montrent aussi que notre méthode surpasse significativement l'approche à base de graphe état de l'art.

Dans cet article, nous proposons trois améliorations simples pour l'apprentissage global d'analyseurs en dépendances par transition de type : un oracle non déterministe, la reprise sur le même exemple après une mise à jour et l'entraînement en configurations sous-optimales. Leur combinaison apporte un gain moyen de 0,2 UAS sur le corpus SPMRL. Nous introduisons également un cadre général permettant la comparaison systématique de ces stratégies et de la plupart des variantes connues. Nous montrons que la littérature n'a étudié que quelques stratégies parmi les nombreuses variations possibles, négligeant ainsi plusieurs pistes d'améliorations potentielles.

Cet article examine l'utilisation du raisonnement analogique dans le contexte de l'apprentissage incrémental. Le problème d'apprentissage sous-jacent développé est le transfert de requêtes formulées en langue naturelle vers des commandes dans un langage de programmation. Nous y explorons deux questions principales : Comment se comporte le raisonnement par analogie dans le contexte de l'apprentissage incrémental ? De quelle manière la séquence d'apprentissage influence-t-elle la

performance globale ? Pour y répondre, nous proposons un protocole expérimental simulant deux utilisateurs et différentes séquences d'apprentissage. Nous montrons que l'ordre dans la séquence d'apprentissage incrémental n'a d'influence notable que sous des conditions spécifiques. Nous constatons également la complémentarité de l'apprentissage incrémental avec l'analogie pour un nombre d'exemples d'apprentissage minimal.

Dans cet article, nous abordons une tâche encore peu explorée, consistant à extraire automatiquement
l'état de l'art d'un domaine scientifique à partir de l'analyse d'articles de ce domaine. Nous la ramenons à deux sous-tâches élémentaires : l'identification de concepts et la reconnaissance de relations entre ces concepts. Une extraction terminologique permet d'identifier les concepts candidats,
qui sont ensuite alignés à des ressources externes. Dans un deuxième temps, nous cherchons à reconnaître et classifier automatiquement les relations sémantiques entre concepts de manière non-supervisée, en nous appuyant sur différentes techniques de clustering et de biclustering. Nous mettons
en oeuvre ces deux étapes dans un corpus extrait de l'archive de l'ACL Anthology. Une analyse manuelle nous a permis de proposer une typologie des relations sémantiques, et de classifier un échantillon d'instances de relations. Les premières évaluations suggèrent l'intérêt du biclustering pour détecter de nouveaux types de relations dans le corpus.

La lisibilité d'un texte dépend fortement de la difficulté des unités lexicales qui le composent. La simplification lexicale vise ainsi à remplacer les termes complexes par des équivalents sémantiques plus simples à comprendre : par exemple, BLEU ('résultat d'un choc') est plus simple que CONTUSION
ou ECCHYMOSE . Il est pour cela nécessaire de disposer de ressources qui listent des synonymes

pour
des sens donnés et les trient par ordre de difficulté. Cet article décrit une méthode pour constituer une ressource de ce type pour le français. Les listes de synonymes sont extraites de BabelNet et de JeuxDeMots, puis triées grâce à un algorithme statistique d'ordonnancement. Les résultats du tri sont
évalués par rapport à 36 listes de synonymes ordonnées manuellement par quarante annotateurs.

Nous évaluons deux modèles sémantiques distributionnels au moyen d'un jeu de données représentant
quatre types de relations lexicales et analysons l'influence des paramètres des deux modèles. Les résultats indiquent que le modèle qui offre les meilleurs résultats dépend des relations ciblées, et que
l'influence des paramètres des deux modèles varie considérablement en fonction de ce facteur. Ils montrent également que ces modèles captent aussi bien la dérivation syntaxique que la synonymie, mais que les configurations qui captent le mieux ces deux types de relations sont très différentes.

Cet article propose de calculer, via Wikipedia, un indice de notoriété pour les entrées du dictionnaire relationnel multilingue de noms propres Prolexbase. Cet indice de notoriété dépend de la langue et participera, d'une part, à la construction d'un module de Prolexbase pour la langue arabe et, d'autre part, à la révision de la notoriété actuellement présente pour les autres langues de la base. Pour calculer
la notoriété, nous utilisons la méthode SAW (précédée du calcul de l'entropie de Shannon) à partir de
cinq valeurs numériques déduites de Wikipedia.

L'extraction de lexiques bilingues à partir de corpus comparables se réalise traditionnellement en

s'appuyant sur deux langues. Des travaux précédents en extraction de lexiques bilingues à partir de corpus parallèles ont démontré que l'utilisation de plus de deux langues peut être utile pour améliorer la qualité des alignements extraits. Nos travaux montrent qu'il est possible d'utiliser la même stratégie pour des corpus comparables. Nous avons défini deux méthodes originales impliquant
des langues pivots et nous les avons évaluées sur quatre langues et deux langues pivots en particulier.
Nos expérimentations ont montré que lorsque l'alignement entre la langue source et la langue pivot est de bonne qualité, l'extraction du lexique en langue cible s'en trouve améliorée.

Inbenta développe un outil de classification non-supervisée hybride qui allie à la fois les statistiques et la puissance de notre lexique inspiré de la Théorie Sens-Texte. Nous présenterons ici le contexte qui a amené à la nécessité de développer un tel outil. Après un rapide état de l'art sur la classification
non-supervisée en TAL, nous décrirons le fonctionnement de notre clustering sémantique.

Nous proposons dans cet article une analyse des résultats de la campagne SemDis 2014 qui proposait
une tâche de substitution lexicale en français. Pour les 300 phrases du jeu de test, des annotateurs ont proposé des substituts à un mot cible, permettant ainsi d'établir un gold standard sur lequel les systèmes participants ont été évalués. Nous cherchons à identifier les principales caractéristiques des
items du jeu de test qui peuvent expliquer les variations de performance pour les humains comme pour
les systèmes, en nous basant sur l'accord inter-annotateurs des premiers et les scores de rappel des

seconds. Nous montrons que si plusieurs caractéristiques communes sont associées aux deux types
de difficulté (rareté du sens dans lequel le mot-cible est employé, fréquence d'emploi du mot-cible), d'autres sont spécifiques aux systèmes (degré de polysémie du mot-cible, complexité syntaxique).

Nous détectons dans des corpus d'avis clients en français des expressions d'opinion ne contenant pas
de marqueur d'opinion explicitement positif ou négatif. Nous procédons pour cela en deux étapes en nous appuyant sur des méthodes existantes : nous identifions ces expressions à l'aide de fenêtres de
mots puis nous les classifions en polarité. Le processus global présente des résultats satisfaisants pour
notre cadre applicatif demandant une haute précision.

La reconnaissance d'entités nommées (REN) pour les langues naturelles telles que l'arabe est une tâche essentielle et difficile. Dans cet article, nous décrivons notre système hybride afin d'améliorer la performance du système de REN et de combler le manque de ressources pour le TAL arabe. Notre
système applique un modèle CRF, un lexique bilingue d'ENs et des règles linguistiques spécifiques à la tâche de reconnaissance d'entités nommées dans les textes arabes. Les résultats empiriques indiquent que notre système surpasse l'état-de l'art de la REN arabe lorsqu'il est appliqué au corpus d'évaluation standard ANERcorp.

L'objectif de cette étude est d'expérimenter l'intégration d'une nouvelle forme d'évaluation dans un correcteur orthographique et grammatical. L'« anti-correcteur » a pour objet de mesurer le taux de réussites orthographiques et grammaticales d'un texte sur certains points jugés difficiles selon la

littérature et une observation d'erreurs en corpus. L'évaluation du niveau d'écriture ne se base plus uniquement sur les erreurs commises, mais également sur les réussites réalisées. Une version bêta de ce nouveau mode d'évaluation positive a été intégré dans le correcteur Cordial. Cet article a pour but de discuter de l'intérêt de ce nouveau rapport à l'orthographe et de présenter quelques premiers éléments d'analyse résultant de l'application de l'anti-correcteur sur un corpus de productions variées en matière de niveau d'écriture et genre discursif.

Dans cet article, nous présentons une méthode pour améliorer la traduction automatique d'un corpus
annoté et porter ses annotations de l'anglais vers une langue cible. Il s'agit d'améliorer la méthode de (Nasiruddin et al., 2015) qui donnait de nombreux segments non traduits, des duplications et des désordres. Nous proposons un processus de pré-traitement du SemCor anglais, pour qu'il soit adapté
au système de traduction automatique statistique utilisé, ainsi qu'un processus de post-traitement pour la sortie. Nous montrons une augmentation de 2,9 points en terme de score F1 sur une tâche de
désambiguïsation lexicale ce qui prouve l'efficacité de notre méthode.

Il est d'usage de mesurer les performances d'un système de résumé automatique en utilisant la métrique ROUGE. Malheureusement, cette métrique n'est pas appropriée pour des approches non supervisées. Nous montrons qu'il est cependant possible d'effectuer une optimisation pour une solution approchée de ROUGE-n en utilisant une fonction objective fondée sur une version pondérée
par document de ROUGE : "document-weighted ROUGE". Cette méthode permet d'obtenir des performances au niveau de l'état de l'art pour des systèmes de résumé automatique pour le français et

l'anglai. Et ceci malgré le fait qu'il n'y ait pas de corrélation entre la métrique ROUGE pondérée au niveau des documents et les jugements humains, contrairement à la métrique ROUGE originale. Ces
résultats suggèrent l'existence en théorie d'une relation d'approximation entre ces deux métriques.

La présente communication s'inscrit dans le cadre du développement d'une grammaire formelle pour
la langue des signes française (LSF). Générer automatiquement des énoncés en LSF implique la définition de certaines règles de production pour synchroniser les différents articulateurs du corps, signes, mouvements, etc. Cet article présente dans sa première partie notre méthodologie pour définir des règles de production à partir d'une étude de corpus. Dans la deuxième partie nous présenterons notre étude qui portera sur deux règles de production pour juxtaposer quelques types de structures en LSF. Nous finissons par une discussion sur la nature et l'apport de notre démarche par rapport aux approches existantes.

Cet article présente les méthodes mises en oeuvre et les résultats obtenus pour la création d'un lexique de formes fléchies de l'alsacien. Les dialectes d'Alsace font partie des langues peu dotées : rares sont les outils et ressources informatisées les concernant. Plusieurs difficultés doivent être prises en compte afin de générer des ressources pour ces langues, généralement liées à la variabilité
en l'absence de norme graphique, et au manque de formes fléchies dans les quelques ressources existantes. Nous avons pour ce faire utilisé plusieurs outils permettant la génération automatique de variantes graphiques et la création de formes fléchies (graphes morphologiques et de flexion d'Unitex).
Les résultats en termes de couverture des formes rencontrées dans des textes ont permis l'évaluation

de la méthode.

Nos travaux portent sur la construction rapide d'outils d'analyse linguistique pour des langues peu dotées en ressources. Dans une précédente contribution, nous avons proposé une méthode pour la construction automatique d'un analyseur morpho-syntaxique via une projection interlingue d'annotations linguistiques à partir de corpus parallèles (méthode fondée sur les réseaux de neurones
récurrents). Nous présentons, dans cet article, une amélioration de notre modèle neuronal, avec la prise
en compte d'informations linguistiques externes pour un annotateur plus complexe. En particulier, nous proposons d'intégrer des annotations morpho-syntaxiques dans notre architecture neuronale pour l'apprentissage non supervisé d'annotateurs sémantiques multilingues à gros grain (annotation en SuperSenses). Nous montrons la validité de notre méthode et sa généricité sur l'italien et le français
et étudions aussi l'impact de la qualité du corpus parallèle sur notre approche (généré par traduction manuelle ou automatique). Nos expériences portent sur la projection d'annotations de l'anglais vers le français et l'italien.

Dans cet article, nous présentons le développement d'un système d'extraction d'expressions-cibles pour l'anglais et sa transposition au français. En complément, nous avons réalisé une étude de l'efficacité des traits en anglais et en français qui tend à montrer qu'il est possible de réaliser un système d'extraction d'expressions-cibles indépendant du domaine. Pour finir, nous proposons une analyse comparative des erreurs commises par nos systèmes en anglais et français et envisageons différentes solutions à ces problèmes.

Nous présentons des travaux récents réalisés autour de MElt, système discriminant d'étiquetage en

parties du discours. MElt met l'accent sur l'exploitation optimale d'informations lexicales externes pour améliorer les performances des étiqueteurs par rapport aux modèles entraînés seulement sur des
corpus annotés. Nous avons entraîné MElt sur plus d'une quarantaine de jeux de données couvrant plus d'une trentaine de langues. Comparé au système état-de-l'art MarMoT, MElt obtient en moyenne
des résultats légèrement moins bons en l'absence de lexique externe, mais meilleurs lorsque de telles
ressources sont disponibles, produisant ainsi des étiqueteurs état-de-l'art pour plusieurs langues.

La reconnaissance d'entités nommées consiste à classer des objets textuels dans des catégories pré-définies telles que les personnes, les lieux et les organisations. Alors que cette tâche suscite de nombreuses études depuis 20 ans, l'application dans des domaines spécialisés reste un défi important.
Nous avons développé un système à base de règles et deux systèmes d'apprentissage pour résoudre
la même tâche : la reconnaissance de noms de produits, de marques, etc., dans le domaine de
la Cosmétique, pour le français. Les systèmes développés peuvent ainsi être comparés dans des conditions idéales. Dans ce papier, nous présentons nos systèmes et nous les comparons.

Cet article présente une expérience d'annotation morphosyntaxique fine du volet serbe du corpus parallèle ParCoLab (corpus serbe-français-anglais). Elle a consisté à enrichir une annotation existante en parties du discours avec des traits morphosyntaxiques fins, afin de préparer une étape ultérieure de parsing. Nous avons comparé trois approches : 1) annotation manuelle ; 2) pré-annotation avec un étiqueteur entraîné sur le croate suivie d'une correction manuelle ; 3) ré-entraînement de l'outil sur un petit échantillon validé du corpus, suivi de l'annotation automatique

et

de la correction manuelle. Le modèle croate maintient une stabilité globale en passant au serbe, mais

les différences entre les deux jeux d'étiquettes exigent des interventions manuelles importantes. Le modèle ré-entraîné sur un échantillon de taille limité (20K tokens) atteint la même exactitude que le modèle existant et le gain de temps observé montre que cette méthode optimise la phase de correction.

En sciences humaines et plus particulièrement en philosophie, l'analyse conceptuelle (AC) est une pratique fondamentale qui permet de décortiquer les propriétés d'un concept. Lors de l'analyse d'un un corpus textuel, le principal défi est l'identification des segments de texte qui expriment le concept.

Parfois, ces segments sont facilement reconnaissables grâce à une unité lexicale attendue, appelée forme canonique. Toutefois, ce n'est pas toujours le cas. Cet article propose une chaîne de traitement

pour la découverte d'un certain nombre de segments périphériques, dits périsegments. Pour illustrer le processus, nous réalisons des expérimentations sur le concept d'« esprit » dans les Collected Papers

de Ch. S. Peirce, en obtenant une précision moyenne supérieure à 83%.

Cet article présente un processus de compilation d'une grammaire de propriétés en une contrainte en extension. Le processus s'insère dans le cadre d'un analyseur syntaxique robuste par résolution d'un problème d'optimisation de contraintes. La grammaire compilée est une énumération de tous les

constituants immédiats uniques de l'espace de recherche. L'intérêt de ce travail encore préliminaire tient principalement dans l'exploration d'une modélisation computationnelle de la langue à base de

Syntaxe par Modèles (MTS, Model-Theoretic Syntax), qui intègre la représentation indifférenciée des énoncés canoniques et non-canoniques. L'objectif plus particulier du travail présenté ici est d'explorer la possibilité de construire l'ensemble des structures candidat-modèles à partir de l'ensemble
des structures syntagmatiques observées sur corpus. Cet article discute notamment le potentiel en matière d'intégration de prédictions probabilistes dans un raisonnement exact pour contribuer à la discrimination entre analyses grammaticales et agrammaticales.

Cet article présente une méthode simple de transfert cross-lingue de dépendances. Nous montrons tout d'abord qu'il est possible d'apprendre un analyseur en dépendances par transition à partir
de données partiellement annotées. Nous proposons ensuite de construire de grands ensembles de données partiellement annotés pour plusieurs langues cibles en projetant les dépendances via les liens d'alignement les plus sûrs. En apprenant des analyseurs pour les langues cibles à partir de ces
données partielles, nous montrons que cette méthode simple obtient des performances qui rivalisent avec celles de méthodes état-de-l'art récentes, tout en ayant un coût algorithmique moindre.

La segmentation d'un texte en rhèses, unités-membres signifiantes de la phrase, permet de fournir des adaptations de celui-ci pour faciliter la lecture aux personnes dyslexiques. Dans cet article, nous proposons une méthode d'identification automatique des rhèses basée sur un apprentissage supervisé
à partir d'un corpus que nous avons annoté. Nous comparons celle-ci à l'identification manuelle ainsi
qu'à l'utilisation d'outils et de concepts proches, tels que la segmentation d'un texte en chunks.

La communication par SMS (Short Message Service), aussi bien que tout autre type de

communication
virtuelle sous forme de textes courts (mails, microblogs, tweets, etc.), présente certaines particularités
spécifiques (syntaxe irrégulière, fusionnement et phonétisation de mots, formes abrégées, etc.). A cause de ces caractéristiques, l'application d'outils en Traitement Automatique du Langage (TAL) rend
difficile l'exploitation d'informations utiles contenues dans des messages bruités. Nous proposons un modèle de normalisation en deux étapes fondé sur une approche symbolique et statistique. La première partie vise à produire une représentation intermédiaire du message SMS par l'application des grammaires locales, tandis que la deuxième utilise un système de traduction automatique à base
de règles pour convertir la représentation intermédiaire vers une forme standard.

À la suite des travaux de Gillick & Favre (2009), beaucoup de travaux portant sur le résumé par extraction se sont appuyés sur une modélisation de cette tâche sous la forme de deux contraintes antagonistes : l'une vise à maximiser la couverture du résumé produit par rapport au contenu des textes d'origine tandis que l'autre représente la limite du résumé en termes de taille. Dans cette approche, la notion de redondance n'est prise en compte que de façon implicite. Dans cet article, nous reprenons le cadre défini par Gillick & Favre (2009) mais nous examinons comment et dans quelle mesure la prise en compte explicite de la similarité sémantique des phrases peut améliorer les
performances d'un système de résumé multi-document. Nous vérifions cet impact par des évaluations
menées sur les corpus DUC 2003 et 2004.

L'objectif de notre travail est d'évaluer l'intérêt d'employer les n-grammes et l'analyse factorielle des

correspondances (AFC) pour comparer les genres textuels dans les études contrastives interlinguistiques. Nous exploitons un corpus bilingue anglais-français constitué de textes originaux comparables.

Le corpus réunit trois genres : les débats parlementaires européens, les éditoriaux de presse et les articles scientifiques. Dans un premier temps, les n-grammes d'une longueur de 2 à 4 mots sont extraits dans chaque langue. Ensuite, pour chaque longueur, les 1 000 n-grammes les plus fréquents
dans chaque langue sont traités par l'AFC pour déterminer quels n-grammes sont particulièrement saillants dans les genres étudiés. Enfin, les n-grammes sont catégorisés manuellement en distinguant
les expressions d'opinion et de certitude, les marqueurs discursifs et les expressions référentielles.

Les résultats montrent que les n-grammes permettent de mettre au jour des caractéristiques typiques
des genres étudiés, de même que des contrastes interlangues intéressants.

L'analyse temporelle des documents cliniques permet d'obtenir des représentations riches des informations contenues dans les dossiers électroniques patient. Cette analyse repose sur l'extraction
d'événements, d'expressions temporelles et des relations entre eux. Dans ce travail, nous considérons
que nous disposons des événements et des expressions temporelles pertinents et nous nous intéressons
aux relations temporelles entre deux événements ou entre un événement et une expression temporelle.

Nous présentons des modèles de classification supervisée pour l'extraction de des relations en français

et en anglais. Les performances obtenues sont comparables dans les deux langues, suggérant ainsi que différents domaines cliniques et différentes langues pourraient être abordés de manière similaire.

Dans cet article, nous proposons une méthode d'appariement de contenus d'actualité multimédias, considérant les exigences à la fois sémantiques et temporelles du besoin d'information. La pertinence
d'une vidéo pour un article de presse est mesurée par deux indices, l'un saisissant la similarité de leurs
contenus, l'autre la cohérence de leurs dates d'édition. Nous présentons également une méthodologie
d'évaluation s'affranchissant des standards comparant les résultats du système à des résultats de référence, en soumettant les paires de documents proposées automatiquement à un panel d'utilisateurs
chargé de juger de leur pertinence.

Pour orienter efficacement les messages reçus par différents canaux de communication, dont l'agent
virtuel (AV), un système de gestion de la relation client doit prendre en compte le besoin d'information de l'usager. En vue d'une tâche de classification par type de besoin d'information, il est utile de pouvoir en amont sélectionner dans les messages des utilisateurs, souvent de mauvaise qualité, les unités textuelles qui seront pertinentes pour représenter ce besoin d'information. Après avoir décrit les spécificités d'un corpus de requêtes d'AV nous expérimentons deux méthodes de sélection de segments informatifs : par extraction et par filtrage. Les résultats sont encourageants, mais des améliorations et une évaluation extrinsèque restent à faire.

Dans cet article nous nous intéressons à la tâche d'extraction de relations sémantiques dans les textes
médicaux et plus particulièrement dans les comptes rendus radiologiques. L'identification de relations
sémantiques est une tâche importante pour plusieurs applications (recherche d'information, génération
de résumé, etc). Nous proposons une approche fondée sur l'utilisation de patrons sémantiques vérifiant
des contraintes dans une base de connaissances.

Nous utilisons des modèles sémantiques distributionnels pour détecter des termes qui évoquent le même cadre sémantique. Dans cet article, nous vérifions si une combinaison de différents modèles permet d'obtenir une précision plus élevée qu'un modèle unique. Nous mettons à l'épreuve plusieurs
méthodes simples pour combiner les mesures de similarité calculées à partir de chaque modèle. Les
résultats indiquent qu'on obtient systématiquement une augmentation de la précision par rapport au meilleur modèle unique en combinant des modèles différents.

Cet article présente une approche visant à évaluer automatiquement la difficulté de dictées en vue de
les intégrer dans une plateforme d'apprentissage de l'orthographe. La particularité de l'exercice de la
dictée est de devoir percevoir du code oral et de le retranscrire via le code écrit. Nous envisageons ce double niveau de difficulté à l'aide de 375 variables mesurant la difficulté de compréhension d'un texte ainsi que les phénomènes orthographiques et grammaticaux complexes qu'il contient. Un

sous-ensemble optimal de ces variables est combiné à l'aide d'un modèle par machines à vecteurs de support (SVM) qui classe correctement 56% des textes. Les variables lexicales basées sur la liste
orthographique de Catach (1984) se révèlent les plus informatives pour le modèle.

Les médias traditionnels sont de plus en plus présents sur les réseaux sociaux, mais ces sources d'informations sont confrontées à d'autres sources dites de réinformation. Ces dernières ont parfois tendance à déformer les informations relayées pour correspondre aux idéologies qu'elles souhaitent défendre, les rendant partiellement ou totalement fausses. Le but de cet article est, d'une part, de présenter un corpus que nous avons constitué à partir de groupes Facebook de ces deux types de médias.
Nous présentons d'autre part quelques expériences de détection automatique des messages issus des médias de réinformation, en étudiant notamment l'influence d'attributs de surface et d'attributs portant plus spécifiquement sur le contenu de ces messages.

Bien que la Traduction Automatique (TA) se soit concentrée jusqu'à présent sur la traduction de textes
écrits et édités, de plus en plus de travaux sont consacrés à la traduction de textes informels et spontanés
(discours et dialogues). Pour traduire de tels textes relevant de l'« oral-écrit », il devient indispensable
de prendre en compte des informations contextuelles, qu'elles soient de nature extra-linguistique (identité du locuteur, interaction entre le locuteur et l'interlocuteur) ou linguistique (coréférence et phénomènes stylistiques propres à la parole). Or l'intégration d'informations contextuelles dans les systèmes de TA reste limitée dans la plupart des systèmes actuels. Dans cet article, nous présentons

et analysons trois expériences d'intégration du contexte dans un système de TA mettant en jeu des formes de contexte et donc des méthodologies différentes: l'adaptation au genre du locuteur, la traduction de pronoms et la génération de « tag questions » anglaises à partir du français.

Les forums de discussion et les réseaux sociaux sont des sources potentielles de différents types d'information, qui ne sont en général pas accessibles par ailleurs. Par exemple, dans les forums de santé, il est possible de trouver les informations sur les habitudes et le mode de vie des personnes. Ces informations sont rarement partagées avec les médecins. Il est donc possible de se fonder sur ces informations pour évaluer les pratiques réelles des patients. Il s'agit cependant d'une source d'information difficile à traiter, essentiellement à cause des spécificités linguistiques qu'elle présente. Si une première étape pour l'exploration des forums consiste à indexer les termes médicaux présents
dans les messages avec des concepts issus de terminologies médicales, cela s'avère extrêmement compliqué car les formulations des patients sont très différentes des terminologies officielles. Nous proposons une méthode permettant de créer et enrichir des lexiques de termes et expressions désignant
une maladie ou un trouble, avec un intérêt particulier pour les troubles de l'humeur. Nous utilisons des ressources existantes ainsi que des méthodes non supervisées. Les ressources construites dans le
cadre du travail nous permettent d'améliorer la détection de messages pertinents.

Nous proposons une nouvelle méthode pour la création automatique de grammaires lexicalisées syntaxico-sémantiques. A l'heure actuelle, la création de grammaire résulte soit d'un travail manuel soit d'un traitement automatisé de corpus arboré. Notre proposition est d'extraire
à partir de données VerbNet une grammaire noyau (formes canoniques des verbes et des groupes nominaux) de l'anglais intégrant une sémantique VerbNet. Notre objectif est de

profiter des larges ressources existantes pour produire un système de génération de texte symbolique de qualité en domaine restreint.

Dans cet article, nous explorons divers traits proposés dans la littérature afin de fournir un détecteur d'entités nommées pour le Français appris automatiquement sur le French Treebank. Nous étudions l'intégration de connaissances en domaine, l'apport de la classification des verbes, la gestion des mots inconnus et l'intégration de traits non locaux. Nous comparons ensuite notre système aux récents réseaux de neurones.

L'émergence des corpus scolaires et la volonté d'outiller ces corpus spécifiques font apparaitre de nouvelles problématiques de recherche pour le traitement automatique des langues (TAL). Nous exposons ici une recherche qui vise le traitement de productions d'apprenants en début d'apprentissage de l'écriture, en vue d'une annotation et d'une exploitation ultérieure. À cette fin, nous proposons d'envisager cette étape comme une tâche d'alignement entre la production de l'apprenant et une normalisation produite manuellement. Ce procédé permet d'augmenter significativement les scores d'identification des formes et lemmes produits et améliore les perspectives d'annotation.

La rareté des ressources numériques pour la langue arabe, telles que les grammaires et corpus, rend
son traitement plus difficile que les autres langues naturelles. A ce jour il n'existe pas une grammaire
formelle à large couverture de l'arabe. Dans ce papier, nous présentons une nouvelle approche qui facilite la description de l'arabe avec le formalisme des grammaires d'arbres adjoints en utilisant une méta-grammaire. Nous exposons les premiers résultats de notre grammaire ainsi que les problèmes rencontrés pour son évaluation.

Dans cet article, nous proposons une classification des déterminants en étudiant leur capacité à introduire de nouveaux référents du discours et l'accessibilité de ces référents. Cette classification se fonde
sur des aspects de logique dynamique (Groenendijk et Stokhof, 1991) dans la tradition montagovienne.
Nous montrons ensuite que ces classes raffinent d'autres classifications plus linguistiques en étudiant
chaque espèce de déterminants une à une. L'analyse de ces propriété est un première étape dans la
définition des quantificateurs généralisés dynamiques nécessaires pour dénoter la sémantique des déterminants.

L'un des objectifs de nos travaux, à terme, est de transformer un corpus de documents médicaux en données structurées pour en faciliter l'exploitation. Ainsi, il est nécessaire non seulement de détecter
les concepts médicaux évoqués, mais aussi d'intégrer un processus capable d'identifier le contexte dans lequel est évoqué chaque concept médical. Dans cet article, nous revenons principalement sur les systèmes par apprentissage supervisé qui ont été proposé pour la détection de l'incertitude et de la négation. Ces dix dernières années, les travaux pour détecter l'incertitude et la négation dans les textes en anglais ont donné des résultats satisfaisants. Cependant, il existe encore une marge de
progression non-négligeable.

Nous proposons dans cet article une méthode semi-supervisée originale pour la création de représentations vectorielles pour des termes (complexes ou non) dans un espace sémantique

pertinent

pour une tâche de normalisation de termes désignant des entités dans un corpus. Notre méthode s'appuie en partie sur une approche de sémantique distributionnelle, celle-ci générant des vecteurs initiaux pour chacun des termes extraits. Ces vecteurs sont alors plongés dans un autre espace vectoriel construit à partir de la structure d'une ontologie. Pour la construction de ce second espace vectoriel ontologique, plusieurs méthodes sont testées et comparées. Le plongement s'effectue par entraînement d'un modèle linéaire. Un calcul de distance (en utilisant la similarité cosinus) est enfin effectué pour déterminer la proximité entre vecteurs de termes et vecteurs de concepts de l'ontologie

servant à la normalisation. La performance de cette méthode a atteint un rang honorable, ouvrant d'encourageantes perspectives.

Cet article a pour but de présenter une démarche généraliste pour l'annotation automatique des lieux

dans l'oral transcrit. Cette annotation est effectuée sur le corpus ESLO (Enquête SocioLinguistique à

Orléans) et suppose une réflexion sur les caractéristiques propres à la désignation d'un lieu à l'oral. Avant d'expliciter la méthode employée pour traiter automatiquement notre corpus, nous présenterons le travail préparatoire de la constitution d'une convention d'annotation et d'un corpus de référence indispensable pour l'évaluation du système.

Dans cet article nous considérons l'apport du Traitement Automatique des Langues (TAL) au problème de la détection automatique de « l'embellissement » (en anglais « spin ») des résultats de recherche dans les publications scientifiques du domaine biomédical. Nous cherchons à identifier les

affirmations inappropriées dans les articles, c'est-à-dire les affirmations où l'effet positif du

traitement étudié est plus grand que celui effectivement prouvé par la recherche. Après une description du problème de point de vue du TAL, nous présentons les pistes de recherche qui nous semblent les plus prometteuses pour automatiser la détection de l'embellissement. Ensuite nous analysons l'état de l'art sur les tâches comparables et présentons les premiers résultats obtenus dans
notre projet avec des méthodes de base (grammaires locales) pour la tâche de l'extraction des entités
spécifiques à notre objectif.

Cet article propose un algorithme efficace (en O(n 4)) pour trouver la catégorie d'un mot manquant dans un énoncé incomplet. Notre travail fait appel à l'algorithme d'unification comme lors de l'apprentissage des grammaires catégorielles et à la programmation dynamique comme dans l'algorithme
Cocke-Younger-Kasami. En utilisant l'interface syntaxique / sémantique des grammaires catégorielles,
ce travail peut être utilisé pour dériver les lectures sémantiques possibles d'un énoncé incomplet. Des
exemples suivis illustrent notre propos.

La reconnaissance et le traitement approprié des expressions polylexicales (EP) constituent un enjeu
pour différentes applications en traitement automatique des langues. Ces expressions sont susceptibles
d'apparaître sous d'autres formes que leur forme canonique, d'où l'intérêt d'étudier leur profil de variabilité. Dans cet article, nous proposons de donner un aperçu de motifs de variation syntaxiques et/ou morphologiques d'après un corpus de 4441 expressions polylexicales verbales (EPV)

annotées

manuellement. L'objectif poursuivi est de générer automatiquement les différentes variantes pour améliorer la performance des techniques de traitement automatique des EPV.

Nous décrivons la partie française des données produites dans le cadre de la campagne multi-lingue

PARSEME sur l'identification d'expressions polylexicales verbales (Savary et al., 2017). Les expressions couvertes pour le français sont les expressions verbales idiomatiques, les verbes intrinsèquement

pronominaux et une généralisation des constructions à verbe support. Ces phénomènes ont été annotés

sur le corpus French-UD (Nivre et al., 2016) et le corpus Sequoia (Candito et Seddah, 2012), soit

un corpus de 22 645 phrases, pour un total de 4 962 expressions annotées. On obtient un ratio d'une

expression annotée tous les 100 tokens environ, avec un fort taux d'expressions discontinues (40%).

Pour déterminer si certaines mesures d'association lexicale fréquemment employées en TAL attribuent

des scores élevés à des n-grammes que le hasard aurait pu produire aussi souvent qu'observé, nous

avons utilisé une extension du test exact de Fisher à des séquences de plus de deux mots. Les analyses

ont porté sur un corpus de quatre millions de mots d'anglais conversationnel extrait du BNC. Les

résultats, basés sur la courbe précision-rappel et sur la précision moyenne, montrent que le LL-simple

est extrêmement efficace. IM3 est plus efficace que les autres mesures basées sur les tests d'hypothèse
et atteint même un niveau de performance presque égal à LL-simple pour les tri-grammes.

Depuis quelques années les réseaux neuronaux se montrent très efficaces dans toutes les tâches de
Traitement Automatique des Langues (TAL). Récemment, une variante de réseau neuronal particulièrement adapté à l'étiquetage de séquences textuelles a été proposée, utilisant des représentations
distributionnelles des étiquettes. Dans cet article, nous reprenons cette variante et nous l'améliorons avec une version profonde. Dans cette version, différentes couches cachées permettent de prendre en
compte séparément les différents types d'informations données en entrée au réseau. Nous évaluons
notre modèle sur les mêmes tâches que la première version de réseau de laquelle nous nous sommes
inspirés. Les résultats montrent que notre variante de réseau neuronal est plus efficace que les autres,
mais aussi qu'elle est plus efficace que tous les autres modèles évalués sur ces tâches, obtenant l'état-de-l'art.

Nous présentons dans cet article une collection de schémas Winograd en français, adaptée de la liste
proposée par Levesque et al. (2012) pour l'anglais. Les schémas Winograd sont des problèmes de résolution d'anaphore conçus pour être IA-complets. Nous montrons que notre collection vérifie deux

propriétés cruciales : elle est robuste vis-à-vis de méthodes statistiques simples ("Google-proof"), tout en étant largement dépourvue d'ambiguïté pour les sujets humains que nous avons testés.

Les essais cliniques sont un élément fondamental pour l'évaluation de nouvelles thérapies ou techniques de diagnostic, de leur sécurité et efficacité. Ils exigent d'avoir un échantillon convenable de
la population. Le défi consiste alors à recruter le nombre suffisant de participants avec des caractéristiques similaires pour garantir que les résultats des essais sont bien contrôlés et dus aux facteurs
étudiés. C'est une tâche difficile, effectuée essentiellement manuellement. Comme les valeurs numériques sont une information très fréquente et importante, nous proposons un système automatique qui
vise leur extraction et normalisation.

Cet article présente un système d'analyse automatique en cadres sémantiques évalué sur un corpus de
textes encyclopédiques d'histoire annotés selon le formalisme FrameNet. L'approche choisie repose sur un modèle intégré d'étiquetage de séquence qui optimise conjointement l'identification des cadres,
la segmentation et l'identification des rôles sémantiques associés. Nous cherchons dans cette étude à
analyser la complexité de la tâche selon plusieurs dimensions. Une analyse détaillée des performances
du système est ainsi proposée, à la fois selon l'angle des paramètres du modèle et de la nature des données.

Notre objectif est l'élaboration d'un système de détection automatique de relations de co-référence le
plus général possible, pour le traitement des anaphores pronominales et les co-références directes. Nous décrivons dans cet article les différentes étapes de traitement des textes dans le système que nous avons développé : (i) l'annotation en traits lexicaux et syntaxiques par le système Macaon ; (ii) le repérage des mentions par un modèle obtenu par apprentissage sur le corpus ANCOR ; (iii) l'annotation sémantique des mentions à partir de deux ressources : le DEM et le LVF ; (iv) l'annotation en co-références par un système à base de règles. Le système est évalué sur le corpus ANCOR.

Les expressions multi-mots jouent un rôle important dans différentes applications du Traitement Automatique de la Langue telles que la traduction automatique et la recherche d'information interlingue. Cet article, d'une part, décrit une approche hybride pour l'acquisition d'un lexique bilingue d'expressions multi-mots à partir d'un corpus parallèle anglais-français, et d'autre part, présente l'impact de l'utilisation d'un lexique bilingue spécialisé d'expressions multi-mots produit par cette approche sur les résultats du système de traduction statistique libre Moses. Nous avons exploré deux métriques basées sur la co-occurrence pour évaluer les liens d'alignement entre les expressions multi-mots des langues source et cible. Les résultats obtenus montrent que la métrique utilisant un dictionnaire bilingue amorce de mots simples améliore aussi bien la qualité de l'alignement d'expressions multi-mots que celle de la traduction.

Nous présentons une nouvelle formalisation de la factivité, la dimension représentant le degré
de croyance qu'une source ? l'auteur ou tout autre agent mentionné dans un texte ? accorde à une éventualité donnée. Nous insistons sur l'aspect dynamique de cette notion ainsi que sur ses interactions
avec la structure discursive. Nous montrons comment une interprétation en termes d'ensembles de

probabilités permet de s'affranchir des principaux problèmes que posait la formalisation utilisée dans
les travaux précédents au calcul d'une factivité cohérente à l'échelle du texte dans sa totalité.

Cet article présente un algorithme implémenté pour l'inférence de patrons d'alternances morphophonologiques entre mots-formes. Il est universel au sens où il permet d'obtenir des classifications
comparables d'une langue à l'autre sans préjuger des types d'alternances. Les patrons constituent une
première étape pour les travaux quantitatifs dans l'approche Mot et Paradigme de la morphologie.

Cet article s'intègre dans un projet collaboratif qui vise à réaliser une analyse longitudinale de
la production universitaire en Géographie. En particulier, nous présentons les premiers résultats
de l'application d'une méthode de détection automatique de métaphores basée sur les modèles de thématiques latentes. Une analyse détaillée permet de mieux comprendre l'impact de certains choix et de réfléchir aux pistes de recherche que nous serons amenés à explorer pour améliorer ces résultats.

Dans cet article, nous modélisons et testons les approches monosémique et polysémique à la multiplicité des sens en morphologie dérivationnelle, en utilisant la sémantique des cadres (Frame semantics)
et XMG. Pour illustrer nos hypothèses et propositions, nous utilisons des exemples de nominalisations
déverbales avec le suffixe -al sur des verbes de changement de possession en anglais (par exemple,
rental, disbursal). Dans notre implementation XMG, nous montrons que le sens sous-spécifié des

affixes ne peut pas toujours être réduit à un simple sens unitaire et qu'en conséquence l'approche polysémique est plus judicieuse que l'approche monosémique. Nous introduisons également des contraintes sur les potentiels référents. Ces contraintes prennent la forme de contraintes de type et spécifient quels arguments de la base verbale sont compatibles avec le référent de la forme dérivée. L'introduction de contraintes de type rend certaines dérivations impossibles en raison d'échecs d'unification entre frames causées par des types incompatibles.

Plusieurs tâches en traitement du langage naturel impliquent de modifier des phrases en conservant au mieux leur sens, comme la reformulation, la compression, la simplification, chacune avec leurs propres données et modèles. Nous introduisons ici une méthode générale s'adressant à tous ces problèmes, utilisant des données plus simples à obtenir : un ensemble de phrases munies d'indicateurs
sur leur style, comme des phrases et le type de sentiment qu'elles expriment. Cette méthode repose sur un modèle d'apprentissage de représentations non supervisé (un auto-encodeur variationnel), puis sur le changement des représentations apprises pour correspondre à un style donné. Le résultat
est évalué qualitativement, puis quantitativement sur le jeu de données de compression de phrases Microsoft, avec des résultats encourageants.

Mesurer la similarité sémantique est à la base de nombreuses applications. Elle joue un rôle important
dans divers domaines tels que la recherche d'information, la traduction automatique, l'extraction d'information ou la détection de plagiat. Dans cet article, nous proposons un système fondé sur le plongement de mots (word embedding). Ce système est destiné à mesurer la similarité sémantique entre des phrases en arabe. L'idée principale est d'exploiter la représentation des mots par des vecteurs

dans un espace multidimensionnel, afin de faciliter leur analyse sémantique et syntaxique. Des pondérations dépendant de la fréquence inverse en documents et de l'étiquetage morpho-syntaxique
sont appliquées sur les phrases examinées, afin d'améliorer l'identification des mots qui sont plus importants dans chaque phrase. La performance de notre système est confirmée par la corrélation de
Pearson entre nos scores de similarité assignés et les jugements humains sur un corpus de référence
de l'état de l'art sur des phrases en arabe.

La tâche de normalisation automatique des messages issus de la communication électronique médiée
requiert une étape préalable consistant à identifier les phénomènes linguistiques. Dans cet article, nous proposons deux typologies pour l'annotation de textes non standard en français, relevant respectivement des niveaux morpho-lexical et morpho-syntaxique. Ces typologies ont été développées en conciliant les typologies existantes et en les faisant évoluer en parallèle d'une annotation manuelle de tweets et de SMS.

Cet article décrit une mesure de similarité sémantique non-supervisée qui repose sur l'introduction d'une matrice de relations entre mots, dans un paradigme de mesure cosinus entre sacs de mots. La
métrique obtenue, apparentée à soft-cosinus, tient compte des relations entre mots qui peuvent être d'ordre lexical ou sémantique selon la matrice considérée. La mise en oeuvre de cette métrique sur la tâche qui consiste à mesurer des similarités sémantiques entre questions posées sur un forum, a remporté la campagne d'évaluation SemEval2017. Si l'approche soumise à la campagne est une combinaison supervisée de différentes mesures non-supervisées, nous présentons dans cet article

en

détail les métriques non-supervisées, qui présentent l'avantage de produire de bons résultats sans nécessiter de ressources spécifiques autres que des données non annotées du domaine considéré.

Ce travail cherche à comprendre pourquoi les performances d'un analyseur morpho-syntaxiques chutent fortement lorsque celui-ci est utilisé sur des données hors domaine. Nous montrons à l'aide d'une expérience jouet que ce comportement peut être dû à un phénomène de masquage des caractéristiques lexicalisées par les caractéristiques non lexicalisées. Nous proposons plusieurs modèles

essayant de réduire cet effet.

Dans cet article, nous proposons une nouvelle méthode pour représenter sous forme vectorielle les sens d'un dictionnaire. Nous utilisons les termes employés dans leur définition en les projetant dans un espace vectoriel, puis en additionnant les vecteurs résultants, avec des pondérations dépendantes

de leur partie du discours et de leur fréquence. Le vecteur de sens résultant est alors utilisé pour trouver des sens reliés, permettant de créer un réseau lexical de manière automatique. Le réseau obtenu est ensuite évalué par rapport au réseau lexical de Word-Net, construit manuellement. Pour cela nous comparons l'impact des différents réseaux sur un système de désambiguïsation lexicale basé sur la mesure de Lesk. L'avantage de notre méthode est qu'elle peut être appliquée à n'importe

quelle langue ne possédant pas un réseau lexical comme celui de Word-Net. Les résultats montrent que notre réseau automatiquement généré permet d'améliorer le score du système de base, atteignant

quasiment la qualité du réseau de Word-Net.

La capture de relations sémantiques entre termes à partir de textes est un moyen privilégié de constituer/alimenter une base de connaissances, ressource indispensable pour l'analyse de textes. Nous proposons et évaluons la combinaison de trois méthodes de production de relations lexico-sémantiques.

La correction des erreurs dans une collection de données est un problème délicat. Elle peut être réalisée manuellement par un expert, ou en utilisant des méthodes de crowdsourcing, ou encore automatiquement au moyen d'algorithmes. Nous présentons ici des méthodes automatiques permettant de détecter les erreurs potentielles « secondaires » induites par les mécanismes automatiques d'inférences de relations, lorsqu'ils s'appuient sur des relations erronées « initiales » détectées manuellement. Des résultats encourageants, mesurés sur le réseau JeuxDeMots, nous invitent à envisager également des stratégies qui permettraient de détecter automatiquement les relations erronées « initiales », ce qui pourrait conduire à une détection automatique de la majorité des erreurs présentes dans le réseau.

Nous avons précédemment montré qu'il est possible de faire produire des annotations syntaxiques de
qualité par des participants à un jeu ayant un but. Nous présentons ici les résultats d'une expérience visant à évaluer leur production sur un corpus plus complexe, en langue de spécialité, en l'occurrence
un corpus de textes scientifiques sur l'ADN. Nous déterminons précisément la complexité de ce corpus, puis nous évaluons les annotations en syntaxe de dépendances produites par les joueurs par
rapport à une référence mise au point par des experts du domaine.

Dans cet article, nous proposons un modèle pour détecter dans les textes générés par des

utilisateurs

(en particulier les tweets), les mots non-standards à corriger. Nous utilisons pour cela des réseaux de neurones convolutifs au niveau des caractères, associés à des "plongements" (embeddings) des mots présents dans le contexte du mot courant. Nous avons utilisé pour l'évaluation trois corpus de référence. Nous avons testé différents modèles qui varient suivant leurs plongements pré-entrainés, leurs configurations et leurs optimisations. Nous avons finalement obtenu une F1-mesure de 0.972 en validation croisée pour la classe des mots non-standards. Cette détection des mots à corriger est l'étape préliminaire pour la normalisation des textes non standards comme les tweets.

Les marqueurs de relation conceptuelle sont un moyen efficace de détecter automatiquement en corpus des Contextes Riches en Connaissances. Dans le domaine de la terminologie ou de l'ingénierie des connaissances, les Contextes Riches en Connaissances peuvent être utiles pour l'élaboration de ressources termino-ontologiques. Bien que la littérature concernant ce sujet soit riche, il n'existe pas de recensement systématique ni d'évaluation à grande échelle des marqueurs de

relation conceptuelle. Pour ces raisons notamment, nous avons constitué une base de marqueurs pour les relations d'hyperonymie, de méronymie, et de cause, en français. Pour chacun de ces marqueurs, son taux de précision est proposé pour des corpus qui varient en fonction du domaine et

du genre textuel.

Récemment, de nouveaux modèles à base de réseaux de neurones récurrents ont été proposés pour

traiter la génération en langage naturel dans des systèmes de dialogue (Wen et al., 2016a). Ces modèles demandent une grande quantité de données d'apprentissage ; or la collecte et l'annotation de

ces données peuvent être laborieuses. Pour répondre à cette problématique, nous nous intéressons ici
à la mise en place d'un protocole d'apprentissage en ligne basé sur un apprentissage par renforcement,
permettant d'améliorer l'utilisation d'un modèle initial appris sur un corpus plus restreint généré
par patrons. Dans cette étude exploratoire, nous proposons une approche basée sur un algorithme de
bandit contre un adversaire, afin d'en étudier l'intérêt et les limites.

Cet article présente trois expériences de détection de mentions dans un corpus de français oral : ANCOR.
Ces expériences utilisent des outils préexistants d'analyse syntaxique du français et des méthodes issues de
travaux sur la co-référence, les anaphores et la détection d'entités nommées. Bien que ces outils ne soient
pas optimisés pour le traitement de l'oral, la qualité de la détection des mentions que nous obtenons est
comparable à l'état de l'art des systèmes conçus pour l'écrit dans d'autres langues. Nous concluons en
proposant des perspectives pour l'amélioration des résultats que nous obtenons et la construction d'un
système end-to-end pour lequel nos expériences peuvent servir de base de travail.

Nous mettons en relief, grâce à une expérimentation avec questionnaires et corpus parallèle, une situation nouvelle en entreprise de rédaction multi-lingue, pour laquelle il n'existe pas de technologie TAL dédiée. Nous suggérons de tirer profit de cette situation inédite de rédacteur traduisant, pour

utiliser l'expertise du rédacteur pendant le processus de traduction et nous préconisons de développer une TA permettant une édition en cours de processus.

L'adaptation au domaine est un verrou scientifique en traduction automatique. Il englobe généralement
l'adaptation de la terminologie et du style, en particulier pour la post-édition humaine dans le cadre d'une traduction assistée par ordinateur. Avec la traduction automatique neuronale, nous étudions une
nouvelle approche d'adaptation au domaine que nous appelons "spécialisation" et qui présente des résultats prometteurs tant dans la vitesse d'apprentissage que dans les scores de traduction. Dans cet
article, nous proposons d'explorer cette approche.

Nous nous intéressons ici à une tâche de détection de concepts dans des textes sans exigence particulière de passage par une phase de détection d'entités avec leurs frontières. Il s'agit donc d'une
tâche de catégorisation de textes multiétiquette, avec des jeux de données annotés au niveau des textes
entiers. Nous faisons l'hypothèse qu'une annotation à un niveau de granularité plus fin, typiquement au niveau de l'énoncé, devrait améliorer la performance d'un détecteur automatique entraîné sur ces
données. Nous examinons cette hypothèse dans le cas de textes courts particuliers : des certificats de
décès où l'on cherche à reconnaître des diagnostics, avec des jeux de données initialement annotés au niveau du certificat entier. Nous constatons qu'une annotation au niveau de la « ligne » améliore effectivement les résultats, mais aussi que le simple fait d'appliquer au niveau de la ligne un

classifieur
entraîné au niveau du texte est déjà une source d'amélioration.

Les revues systématiques de la littérature dans le domaine biomédical reposent essentiellement sur le travail bibliographique manuel d'experts. Nous évaluons les performances de la classification supervisée pour la découverte automatique d'articles à l'aide de plusieurs définitions des critères d'inclusion. Nous appliquons un modèle de regression logistique sur deux corpus issus de revues systématiques conduites dans le domaine du traitement automatique de la langue et de l'efficacité des
médicaments. La classification offre une aire sous la courbe moyenne (AUC) de 0.769 si le classifieur
est contruit à partir des jugements experts portés sur les titres et résumés des articles, et de 0.835 si on
utilise les jugements portés sur le texte intégral. Ces résultats indiquent l'importance des jugements portés dès le début du processus de sélection pour développer un classifieur efficace pour accélérer l'élaboration des revues systématiques à l'aide d'un algorithme de classification standard.

Dans cet article, nous présentons un processus d'identification automatique de l'origine dialectale pour la langue arabe de textes écrits en caractères arabes ou en écriture latine (arabizi). Nous décrivons
le processus d'annotation des ressources construites et du système de translittération adopté. Deux approches d'identification de la langue sont comparées : la première est linguistique et exploite des dictionnaires, la seconde est statistique et repose sur des méthodes traditionnelles d'apprentissage automatique (n-grammes). L'évaluation de ces approches montre que la méthode linguistique donne
des résultats satisfaisants, sans être dépendante des corpus d'apprentissage.

Cet article présente un travail exploratoire sur l'ajout automatique de disfluences, c'est-à-dire de pauses, de répétitions et de révisions, dans les énoncés en entrée d'un système de synthèse de la parole. L'objectif est de conférer aux signaux ainsi synthétisés un caractère plus spontané et expressif.

Pour cela, nous présentons une formalisation novatrice du processus de production de disfluences à travers un mécanisme de composition de ces disfluences. Cette formalisation se distingue notamment des approches visant la détection ou le nettoyage de disfluences dans des transcriptions, ou de celles en synthèse de la parole qui ne s'intéressent qu'au seul ajout de pauses. Nous présentons une première implémentation de notre processus fondée sur des champs aléatoires conditionnels et des modèles de langage, puis conduisons des évaluations objectives et perceptives. Celles-ci nous permettent de conclure à la fonctionnalité de notre proposition et d'en discuter les pistes principales d'amélioration.

Lorsqu'ils sont traduits depuis une langue à morphologie riche vers l'anglais, les mots-formes sources contiennent des marques d'informations grammaticales pouvant être jugées redondantes par rapport à l'anglais, causant une variabilité formelle qui nuit à l'estimation des modèles probabilistes. Un moyen bien documenté pour atténuer ce problème consiste à supprimer l'information non pertinente de la source en la normalisant. Ce pré-traitement est généralement effectué de manière déterministe, à

l'aide de règles produites manuellement. Une telle normalisation est, par essence, sous-optimale et doit être adaptée pour chaque paire de langues. Nous présentons, dans cet article, une méthode simple
pour rechercher automatiquement une normalisation optimale de la morphologie source par rapport à
la langue cible et montrons que celle-ci peut améliorer la traduction automatique.

Cet article propose une architecture neuronale pour un modèle de langue à vocabulaire ouvert. Les représentations continues des mots sont calculées à la volée à partir des caractères les composant, grâce à une couche convolutionnelle suivie d'une couche de regroupement (pooling). Cela permet au modèle de représenter n'importe quel mot, qu'il fasse partie du contexte ou soit évalué pour la prédiction. La fonction objectif est dérivée de l'estimation contrastive bruitée (Noise Contrastive Estimation, ou NCE), calculable dans notre cas sans vocabulaire. Nous évaluons la capacité de notre modèle à construire des représentations continues de mots inconnus sur la tâche de traduction
automatique IWSLT-2016, de l'Anglais vers le Tchèque, en ré-évaluant les N meilleures hypothèses (N-best re-ranking). Les résultats expérimentaux permettent des gains jusqu'à 0,7 point BLEU. Ils montrent aussi la difficulté d'utiliser des représentations dérivées des caractères pour la prédiction.

Cette recherche a pour principal objectif d'évaluer l'utilité de prendre en compte des mesures totalement automatiques de la compétence phraséologique pour estimer la qualité de textes d'apprenants
de l'anglais langue étrangère. Les analyses, menées sur plus de 1000 copies d'examen du First Certificate in English, librement mises à disposition par Yannakoudakis et coll., confirment que l'approche qui consiste à assigner aux bi-grammes et aux tri-grammes de mots présents dans un texte

des scores d'association collocationnelle calculés sur la base d'un grand corpus de référence natif est
particulièrement efficace. Si les indices extraits des tri-grammes sont moins efficaces que ceux extraits
des bi-grammes, ils apportent une contribution utile à ces derniers. Les analyses soulignent aussi les
bénéfices apportés par un emploi simultané de plusieurs mesures d'association collocationnelle.

Cet article présente un système original de traduction de documents numérisés en arabe. Deux modules sont cascadés : un système de reconnaissance optique de caractères (OCR) en arabe et un
système de traduction automatique (TA) arabe-français. Le couplage OCR-TA a été peu abordé dans la littérature et l'originalité de cette étude consiste à proposer un couplage étroit entre OCR et TA ainsi qu'un traitement spécifique des mots hors vocabulaire (MHV) engendrés par les erreurs d'OCRisation. Le couplage OCR-TA par treillis et notre traitement des MHV par remplacement selon une mesure composite qui prend en compte forme de surface et contexte du mot, permettent une amélioration significative des performances de traduction. Les expérimentations sont réalisés sur un corpus de journaux numérisés en arabe et permettent d'obtenir des améliorations en score BLEU de 3,73 et 5,5 sur les corpus de développement et de test respectivement.

Nous présentons de nouvelles instanciations de trois corpus arborés en constituants du français, où certains phénomènes syntaxiques à l'origine de dépendances à longue distance sont représentés directement à l'aide de constituants discontinus. Les arbres obtenus relèvent de formalismes grammaticaux
légèrement sensibles au contexte (LCFRS). Nous montrons ensuite qu'il est possible d'analyser automatiquement de telles structures de manière efficace à condition de s'appuyer sur une méthode

d'inférence approximative. Pour cela, nous présentons un analyseur syntaxique par transitions, qui réalise également l'analyse morphologique et l'étiquetage fonctionnel des mots de la phrase. Enfin, nos expériences montrent que la rareté des phénomènes concernés dans les données françaises pose
des difficultés pour l'apprentissage et l'évaluation des structures discontinues.

Dans cet article, nous nous intéressons à un nouveau problème, appelé plongement de thésaurus, consistant à transformer un thésaurus distributionnel en une représentation dense de mots. Nous proposons de traiter ce problème par une méthode fondée sur l'association d'un plongement de graphe
et de l'injection de relations dans des représentations denses. Nous avons appliqué et évalué cette méthode pour un large ensemble de noms en anglais et montré que les représentations denses produites
obtiennent de meilleures performances, selon une évaluation intrinsèque, que les représentations denses construites selon les méthodes de l'état de l'art sur le même corpus. Nous illustrons aussi l'intérêt de la méthode développée pour améliorer les représentations denses existantes à la fois de façon endogène et exogène.

Dans cet article nous proposons une nouvelle méthode pour le calcul de word embeddings en utilisant
des projections aléatoires. Notre approche est telle que des méthodes de pondération comme positive
pointwise mutual information (PPMI) peuvent être appliquées après la construction de notre modèle véctoriel, i.e., après avoir déjà réduit la dimensionalité. Les word embeddings peuvent alors être tranferés d'une manière efficace et avec une performance de calcul supérieure vers des espaces sémantiquement discriminants. De plus, l'approche se distingue par une mise-à-jour facilitée de

l'espace vectoriel et une interopérabilité augmentée. Nous évaluons notre méthode par rapport à plusieurs tâches sémantiques. Nous montrons qu'elle donne des résultats comparables à ceux des meilleures méthodes appliquées à des corpus monolingues.

L'intégration de la notion de similarité sémantique entre les unités lexicales est essentielle dans différentes applications de Traitement Automatique des Langues (TAL). De ce fait, elle a reçu un intérêt considérable qui a eu comme conséquence le développement d'une vaste gamme d'approches
pour en déterminer une mesure. Ainsi, plusieurs types de mesures de similarité existent, elles utilisent
différentes représentations obtenues à partir d'informations soit dans des ressources lexicales, soit dans de gros corpus de données ou bien dans les deux. Dans cet article, nous nous intéressons à la création de signatures sémantiques décrivant des représentations vectorielles de mots à partir du réseau lexical JeuxDeMots (JDM). L'évaluation de ces signatures est réalisée sur deux tâches différentes : mesures de similarité sémantique et substitution lexicale. Les résultats obtenus sont très
satisfaisants et surpassent, dans certains cas, les performances des systèmes de l'état de l'art.

Nous présentons ici les résultats d'une expérience menée sur l'annotation en parties du discours d'un
corpus d'une langue régionale encore peu dotée, l'alsacien, via une plateforme de myriadisation (crowdsourcing) bénévole développée spécifiquement à cette fin : Bisame 1 . La plateforme, mise en
ligne en mai 2016, nous a permis de recueillir 15 846 annotations grâce à 42 participants. L'évaluation
des annotations, réalisée sur un corpus de référence, montre que la F-mesure des annotations

volontaires est de 0, 93. Le tagger entraîné sur le corpus annoté atteint lui 82 % d'exactitude. Il s'agit du
premier tagger spécifique à l'alsacien. Cette méthode de développement de ressources langagières est
donc efficace et prometteuse pour certaines langues peu dotées, dont un nombre suffisant de locuteurs
est connecté et actif sur le Web. Le code de la plateforme, le corpus annoté et le tagger sont librement
disponibles.

Cet article s'interroge sur les modalités de participation citoyenne aux recherches en TALN, à la lumière des projets actuels en sciences citoyennes mais aussi d'études menées sur le sujet en histoire des sciences. Il vise à montrer comment une science participative est déjà en marche en TALN, à interroger ses modalités et également à en circonscrire les limites.

Les ressources lexicales électroniques ne contiennent quasiment jamais d'informations étymologiques.
De telles informations, convenablement formalisées, permettraient pourtant de développer des outils automatiques au service de la linguistique historique et comparative, ainsi que d'améliorer significativement le traitement automatique de langues anciennes. Nous décrivons ici le processus que nous
avons mis en oeuvre pour extraire des données étymologiques à partir des notices étymologiques du
wiktionary, rédigées en anglais. Nous avons ainsi produit une base multi-lingue de près d'un million de lexèmes et une base de plus d'un demi-million de relations étymologiques entre lexèmes.

La désambiguïsation d'entités (ou liaison d'entités), qui consiste à relier des mentions d'entités d'un texte à des entités d'une base de connaissance, est un problème qui se pose, entre autre, pour le peuplement automatique de bases de connaissances à partir de textes. Une difficulté de cette tâche est
la résolution d'ambiguïtés car les systèmes ont à choisir parmi un nombre important de candidats. Cet
article propose une nouvelle approche fondée sur l'apprentissage joint de représentations distribuées
des mots et des entités dans le même espace, ce qui permet d'établir un modèle robuste pour la comparaison entre le contexte local de la mention d'entité et les entités candidates.

Nous faisons l'hypothèse que les mots techniques inconnus dotés d'une structure interne (mots affixés ou composés) peuvent fournir des indices linguistiques à un locuteur, ce qui peut l'aider à analyser et à comprendre ces mots. Afin de tester notre hypothèse, nous proposons de travailler sur un
ensemble de mots techniques provenant du domaine médical. Un grand ensemble de mots techniques
est annoté par cinq annotateurs. Nous effectuons deux types d'analyses : l'analyse de l'évolution des mots compréhensibles et incompréhensibles (de manière générale et en fonction de certains suffixes) et l'analyse des clusters avec ces mots créés par apprentissage non-supervisé, sur la base des descripteurs linguistiques et extra-linguistiques. Nos résultats indiquent que, selon la sensibilité linguistique des annotateurs, les mots techniques peuvent devenir décodables et compréhensibles. Quant aux clusters, le contenu de certains reflète la difficulté des mots qui les composent et montre également la progression des annotateurs dans leur compréhension. La ressource construite est disponible pour la recherche : http://natalia.grabar.free.fr/rated-lexicon.html .

Cet article présente l'édition 2018 de la campagne d'évaluation DEFT (Défi Fouille de Textes). A partir d'un corpus de tweets, quatre tâches ont été proposées : identifier les tweets sur la thématique
des transports, puis parmi ces derniers, identifier la polarité (négatif, neutre, positif, mixte), identifier les marqueurs de sentiment et la cible, et enfin, annoter complètement chaque tweet en source et cible
des sentiments exprimés. Douze équipes ont participé, majoritairement sur les deux premières tâches.
Sur l'identification de la thématique des transports, la micro F-mesure varie de 0,827 à 0,908. Sur l'identification de la polarité globale, la micro F-mesure varie de 0,381 à 0,823.

Ce papier décrit la participation d'EDF R&D à la campagne d'évaluation DEFT 2018. Notre équipe a participé aux deux premières tâches : classification des tweets en transport/non-transport (Tâche T1)
et détection de la polarité globale des tweets (Tâche T2). Nous avons utilisé 3 méthodes différentes s'appuyant sur Word2Vec, CNN et LSTM. Aucune donnée supplémentaire, autre que les données d'apprentissage, n'a été utilisée. Notre équipe obtient des résultats très corrects et se classe 1 ère équipe
non académique. Les méthodes proposées sont facilement transposables à d'autres tâches de classification de textes courts et peuvent intéresser plusieurs entités du groupe EDF.

Dans cet article, nous présentons notre contribution au Défi Fouille de Textes 2018 au travers de trois méthodes originales pour la classification thématique et la détection de polarité dans des tweets
en français. Nous y avons ajouté un système de vote. Notre première méthode est fondée sur des lexiques (mots et emojis), les n-grammes de caractères et un classificateur à vaste marge (ou

SVM).

tandis que les deux autres sont des méthodes endogènes fondées sur l'extraction de caractéristiques

au grain caractères : un modèle à mémoire à court-terme persistante (ou BiLSTM pour Bidirectionnal

Long Short-Term Memory) et perceptron multi-couche d'une part et un modèle de séquences de caractères fermées fréquentes et classificateur SVM d'autre part. Le BiLSTM a produit de loin les meilleurs résultats puisqu'il a obtenu la première place sur la tâche 1, classification binaire de tweets selon qu'ils traitent ou non des transports, et la troisième place sur la tâche 2, classification de la polarité en 4 classes. Ce résultat est d'autant plus intéressant que la méthode proposée est faiblement

paramétrique, totalement endogène et qu'elle n'implique aucun pré-traitement.

Nous présentons le système utilisé par l'équipe Melodi/Synapse Développement dans la compétition

DEFT2018 portant sur la classification de thématique ou de sentiments de tweets en français. On propose un système unique pour les deux approches qui combine concaténativement deux méthodes

d'embedding et trois modèles de représentation séquence. Le système se classe 1/13 en analyse de

sentiments et 4/13 en classification thématique.

Cet article décrit les systèmes développés par l'équipe LinkMedia de l'IRISA pour la campagne d'évaluation DeFT 2018 portant sur l'analyse d'opinion dans des tweets en français. L'équipe a participé à 3 des 4 tâches de la campagne : (i) classification des tweets selon s'ils concernent les transports ou non, (ii) classification des tweets selon leur polarité et (iii) annotation des marqueurs

d'opinion et de l'objet à propos duquel est exprimée l'opinion. Nous avons utilisé un algorithme de boosting d'arbres de décision et des réseaux de neurones récurrents (RNN) pour traiter les tâches 1 et

2. Pour la tâche 3 nous avons expérimenté l'utilisation de réseaux de neurones récurrents associés à des CRF. Ces approches donnent des résultats proches, avec un léger avantage aux RNN, et ont permis d'être parmi les premiers classés pour chacune des tâches.

Dans le cadre de l'atelier DEFT 2018 nous nous sommes intéressés à la classification de microblogs

(ici, des tweets) rédigés en français. Ici, nous proposons une méthode se basant sur un réseau hiérarchique de neurones récurrent avec attention. La spécificité de notre architecture est de prendre

en compte ?via un mechanisme d'attention et de portes? les hashtags et les mentions directes (e.g., @user), spécifiques aux microblogs. Notre modèle a obtenu de très bon résultats sur la première tâche

et des résultats compétitifs sur la seconde.

Cet article décrit les systèmes de l'équipe Eloquant pour la catégorisation de tweets en français dans

les tâches 1 (détection de la thématique transports en commun) et 2 (détection de la polarité globale)

du DEFT 2018. Nos systèmes reposent sur un enrichissement sémantique, l'apprentissage automatique et, pour la tâche 1 une approche symbolique. Nous avons effectué deux runs pour chacune des tâches. Nos meilleures F-mesures (0.897 pour la tâche 1 et 0.800 pour la tâche 2) sont

au-dessus de la moyenne globale pour chaque tâche, et nous placent dans les 30% supérieurs de

tous

les runs pour la tâche 2.

Dans ce papier, nous décrivons les systèmes développés au LSE pour le DEFT 2018 sur les tâches 1

et 2 qui consistent à classifier des tweets. La première tâche consiste à déterminer si un message concerne les transports ou non. La deuxième, consiste à classifier les tweets selon leur polarité globale. Pour les deux tâches nous avons développé des systèmes basés sur des réseaux de neurones

convolutifs (CNN) et récurrents (LSTM, BLSTM et GRU). Chaque mot d'un tweet donné est représenté par un vecteur dense appris à partir des données relativement proches de celles de la compétition. Le score final officiel est de 0.891 pour la tâche 1 et de 0.781 pour la tâche 2.

Nous présentons la participation de Syllabs à la tâche de classification de tweets dans le domaine du

transport lors de DEFT 2018. Pour cette première participation à une campagne DEFT, nous avons choisi de tester plusieurs algorithmes de classification état de l'art. Après une étape de prétraitement

commune à l'ensemble des algorithmes, nous effectuons un apprentissage sur le seul contenu des tweets. Les résultats étant somme toute assez proches, nous effectuons un vote majoritaire sur les trois algorithmes ayant obtenus les meilleurs résultats.

Dans ce papier, nous décrivons notre participation au défi d'analyse de texte DEFT 2018. Nous avons

participé à deux tâches : (i) classification transport/non-transport et (ii) analyse de polarité globale des tweets : positifs, negatifs, neutres et mixtes. Nous avons exploité un réseau de neurone basé

sur

un perceptron multicouche mais utilisant une seule couche cachée.

Cet article présente une méthode permettant de collecter sur le web des informations complémentaires à une information prédéfinie, afin de remplir une base de connaissances. Notre méthode utilise des patrons lexico-syntaxiques, servant à la fois de requêtes de recherche et de patrons d'extraction permettant l'analyse de documents non structurés. Pour ce faire, il nous a fallu définir au préalable les critères pertinents issus des analyses dans l'objectif de faciliter la découverte de nouvelles valeurs.

La fouille d'opinion est une activité essentielle pour la veille économique, facilitée par les réseaux sociaux et forums dédiés. L'analyse repose généralement sur des lexiques de sentiments. Pourtant, certaines opinions sont exprimées au moyen d'inférences. Dans cet article, nous proposons une classification des inférences utilisées en chinois dans des commentaires touristiques, à des fins de fouille d'opinion, selon trois niveaux d'analyse (réalisation sémantique, modalité de réalisation, et mode de production). Nous démontrons l'intérêt d'analyser les différents types d'inférence pour déterminer la polarité des opinions exprimées en corpus. Nous présentons également de premiers résultats fondés sur des plongements lexicaux.

Notre étude s'inscrit dans le cadre d'une thèse ayant pour but d'exploiter les modèles distributionnels

pour décrire sémantiquement des classes de mots définies selon des critères morphologiques. Nous

utilisons des indices morphologiques et formels fournis par une base lexicale pour cibler les noms agentifs déverbaux construits par suffixation en -eur. Nous montrons qu'il est possible de constituer un représentant prototypique de la classe sémantique des noms agentifs en -eur dans les modèles

distributionnels. L'étude de ce représentant met en évidence que l'information sémantique véhiculée par le suffixe varie en fonction du corpus d'étude et du degré de lexicalisation des dérivés.

Cet article explore la construction de représentations formelles d'énoncés en langue naturelle. Le passage d'un langage naturel à une représentation logique est réalisé avec un formalisme grammatical,
reliant l'analyse syntaxique de l'énoncé à une représentation sémantique. Nous ciblons l'aspect comportemental des cahiers des charges pour les systèmes cyber-physiques, c'est-à-dire tout type de
systèmes dans lesquels des composants logiciels interagissent étroitement avec un environnement physique. Dans ce cadre, l'enjeu serait d'apporter une aide au concepteur. Il s'agit de permettre de simuler et vérifier, par des méthodes automatiques ou assistées, des cahiers des charges "systèmes" exprimés en langue naturelle. Cet article présente des solutions existantes qui pourraient être combinées
en vue de la résolution de la problématique exposée.

Les systèmes de résumé automatique de textes (SRAT) consistent à produire une représentation condensée et pertinente à partir d'un ou de plusieurs documents textuels. La majorité des SRAT sont
basés sur des approches extractives. La tendance actuelle consiste à s'orienter vers les approches abstractives. Dans ce contexte, le résumé guidé défini par la campagne d'évaluation internationale TAC (Text Analysis Conference) en 2010, vise à encourager la recherche sur ce type d'approche, en
se basant sur des techniques d'analyse en profondeur de textes. Dans ce papier, nous nous penchons
sur le résumé automatique guidé de textes. Dans un premier temps, nous définissons les différentes

caractéristiques et contraintes liées à cette tâche. Ensuite, nous dressons un état de l'art des principaux
systèmes existants en mettant l'accent sur les travaux les plus récents, et en les classifiant selon les approches adoptées, les techniques utilisées, et leurs évaluations sur des corpus de références. Enfin,
nous proposons les grandes étapes d'une méthode spécifique devant permettre le développement d'un
nouveau type de systèmes de résumé guidé.

L'article présente une étude des descripteurs linguistiques pour la caractérisation d'un texte selon son
registre de langue (familier, courant, soutenu). Cette étude a pour but de poser un premier jalon pour
des tâches futures sur le sujet (classification, extraction de motifs discriminants). À partir d'un état de
l'art mené sur la notion de registre dans la littérature linguistique et sociolinguistique, nous avons identifié une liste de 72 descripteurs pertinents. Dans cet article, nous présentons les 30 premiers que
nous avons pu valider sur un corpus de textes français de registres distincts.

Les relations de traduction, qui distinguent la traduction littérale d'autres procédés, constituent un sujet d'étude important pour les traducteurs humains (Chuquet et Paillard, 1989). Or les traitements automatiques fondés sur des relations entre langues, tels que la traduction automatique ou la méthode
de génération de paraphrases par équivalence de traduction, ne les ont pas exploitées explicitement

jusqu'à présent. Dans ce travail, nous présentons une catégorisation des relations de traduction et nous
les annotons dans un corpus parallèle multilingue (anglais, français, chinois) de présentations orales,
les TED Talks. Notre objectif à plus long terme sera d'en faire la détection de manière automatique afin de pouvoir les intégrer comme caractéristiques importantes pour la recherche de segments monolingues en relation d'équivalence (paraphrases) ou d'implication. Le corpus annoté résultant de notre travail sera mis à disposition de la communauté.

Cet article correspond à un état de l'art sur le thème de l'annotation automatique d'images d'observation de la terre pour la détection de la déforestation. Nous nous intéressons aux différents challenges
que recouvre le domaine et nous présentons les méthodes de l'état de l'art puis les pistes de recherche
que nous envisageons.

Les influenceurs ont la capacité d'avoir un impact sur d'autres individus lorsqu'ils interagissent avec eux. Détecter les influenceurs permet d'identifier les quelques individus à cibler pour toucher largement un réseau. Il est possible d'analyser les interactions dans un média social du point de vue de leur structure ou de leur contenu. Dans nos travaux de thèse, nous abordons ces deux aspects. Nous présentons d'abord une évaluation de différentes mesures de centralité sur la structure d'interactions extraites de Twitter puis nous analysons l'impact de la taille du graphe de suivi sur la performance de mesures de centralité. Nous abordons l'aspect linguistique pour identifier le changement d'avis comme un effet de l'influence depuis les messages d'un forum.

Cet article présente et analyse les premiers résultats obtenus par notre laboratoire pour la

construction d'un modèle de résolution des coréférences en français à l'aide de techniques de classifications parmi lesquelles les arbres de décision et les séparateurs à vaste marge. Ce système a été entraîné sur le corpus ANCOR et s'inspire de travaux antérieurs réalisés au laboratoire LATTICE (système CROC). Nous présentons les expérimentations que nous avons menées pour améliorer le système en passant par des classifieurs spécifiques à chaque type de situation interactive, puis chaque type de relation de coréférence.

Notre travail traite de la simplification automatique de textes. Ce type d'application vise à rendre des contenus difficiles à comprendre plus lisibles. À partir de trois corpus comparables du domaine médical, d'un lexique existant et d'une terminologie du domaine, nous procédons à des analyses et à des modifications en vue de la simplification lexicale de textes médicaux. L'alignement manuel des phrases provenant de ces corpus comparables fournit des données de référence et permet d'analyser les procédés de simplification mis en place. La substitution lexicale avec la ressource existante permet d'effectuer de premiers tests de simplification lexicale et indique que des ressources plus spécifiques sont nécessaires pour traiter les textes médicaux. L'évaluation des substitutions est effectuée avec trois critères : grammaticalité, simplification et sémantique. Elle indique que la grammaticalité est plutôt bien sauvegardée, alors que la sémantique et la simplicité sont plus difficiles à gérer lors des substitutions avec ce type de méthodes.

Cet article décrit le développement du premier corpus syntaxiquement annoté de breton. Le corpus fait

partie du projet «Universal Dependencies». Dans cet article, nous décrivons la préparation du corpus,
certaines constructions spécifiques au breton qui avaient besoin d'un traitement spécial et nous donnons
des résultats de l'analyse syntaxique de breton par un nombre d'analyseurs syntaxiques.

La détection de frontières de phrase est généralement considéré comme un problème résolu. Cependant, les outils performant sur des textes en domaine général, ne le sont pas forcement sur des
domaines spécialisés, ce qui peut engendrer des dégradations de performance des outils intervenant
en aval dans une chaîne de traitement automatique s'appuyant sur des textes découpés en phrases.
Dans cet article, nous évaluons 5 outils de segmentation en phrase sur 3 corpus issus de différent domaines. Nous ré-entrainerons l'un de ces outils sur un corpus de spécialité pour étudier l'adaptation
en domaine. Notamment, nous utilisons un nouveau corpus biomédical annoté spécifiquement pour cette tâche. La detection de frontières de phrase à l'aide d'un modèle OpenNLP entraîné sur un corpus
clinique offre une F-mesure de .73, contre .66 pour la version standard de l'outil.

Nous nous intéressons, dans cet article, à la détection d'opinions dans la langue arabe. Ces dernières
années, l'utilisation de l'apprentissage profond a amélioré des performances de nombreux systèmes automatiques dans une grande variété de domaines (analyse d'images, reconnaissance de la parole,

traduction automatique, . . .) et également celui de l'analyse d'opinions en anglais. Ainsi, nous avons étudié l'apport de deux architectures (CNN et LSTM) dans notre cadre spécifique. Nous avons également testé et comparé plusieurs types de représentations continues de mots (embeddings) disponibles en langue arabe, qui ont permis d'obtenir de bons résultats. Nous avons analysé les erreurs
de notre système et la pertinence de ces embeddings. Cette analyse mène à plusieurs perspectives intéressantes de travail, au sujet notamment de la constitution automatique de ressources expert et d'une construction pertinente des embeddings spécifiques à la tâche d'analyse d'opinions.

Cet article présente des méthodes permettant l'évaluation de la satisfaction client à partir de très vastes
corpus de conversation de type "chat" entre des clients et des opérateurs. Extraire des connaissances
dans ce contexte demeure un défi pour les méthodes de traitement automatique des langues de par la dimension interactive et les propriétés de ce nouveau type de langage à l'intersection du langage écrit et parlé. Nous présentons une étude utilisant des réponses à des sondages utilisateurs comme supervision faible permettant de prédire la satisfaction des usagers d'un service en ligne d'assistance
technique et commerciale.

Le traitement à posteriori de transcriptions OCR cherche à détecter les erreurs dans les sorties d'OCR
pour tenter de les corriger, deux tâches évaluées par la compétition ICDAR-2017 Post-OCR Text Correction. Nous présenterons dans ce papier un système de détection d'erreurs basé sur un modèle
à réseaux récurrents combinant une analyse du texte au niveau des mots et des caractères en deux

temps. Ce système a été classé second dans trois catégories évaluées parmi 11 candidats lors de la
compétition.

Cet article présente une tâche du benchmarking de la reconnaissance de l'entité nommée (REN) pour le français. Nous entrainons et évaluons plusieurs algorithmes d'étiquetage de séquence, et nous améliorons les résultats de REN avec une approche fondée sur l'utilisation de l'apprentissage semi-supervisé et du reclassement. Nous obtenons jusqu'à 77.95%, améliorant ainsi le résultat de plus de 34 points par rapport du résultat de base du modèle.

Cet article traite des analyses d'erreurs quantitatives et qualitatives sur les résultats de l'analyse syntaxique des constituants pour le français. Pour cela, nous étendons l'approche de Kummerfeld et al. (2012) pour français, et nous présentons les détails de l'analyse. Nous entraînons les systèmes
d'analyse syntaxique statistiques et neuraux avec le corpus arboré pour français, et nous évaluons les
résultats d'analyse. Le corpus arboré pour le français fournit des étiquettes syntagmatiques à grain fin,
et les caractéristiques grammaticales du corpus affectent des erreurs d'analyse syntaxique.

Dans une grammaire formelle, le lien entre l'information sémantique et sa structure syntaxique correspondante peut être établi en utilisant une interface syntaxe/sémantique qui permettra la construction du sens de la phrase. L'étiquetage de rôles sémantiques aide à réaliser cette tâche en associant automatiquement des rôles sémantiques à chaque argument du prédicat d'une phrase. Dans ce papier, nous présentons une nouvelle approche qui permet la construction d'une telle interface pour une grammaire d'arbres adjoints de l'arabe. Cette grammaire a été générée semi

automatiquement à partir d'une méta-grammaire. Nous détaillons le processus d'interfaçage entre le niveau syntaxique et le niveau sémantique moyennant la sémantique des cadres et comment avons-nous procédé à l'étiquetage de rôles sémantiques en utilisant la ressource lexicale ArabicVerbNet.

L'objectif de cet article est de présenter la construction d'un système d'analyse de sentiments dans le domaine des microblogs financiers en anglais. Le but de notre travail est de construire un classifieur pour la prédiction de sentiments chez les investisseurs financiers sur les plateformes de microblogs telles que StockTwits et Twitter. Notre contribution montre qu'il est possible de mener une analyse fine des sentiments. Après extraction des entités financières et leurs contextes, le système attribue des scores en valeurs continues. Il repose sur une approche par réseaux profonds pour la méthode de classification. Les résultats montrent un F1-score de 0.85 (2 classes) et une valeur de similarité cosinus de 0.62.

Nous proposons trois nouvelles méthodes pour construire et optimiser des plongements de mots pour le français. Nous utilisons les résultats de l'étiquetage morpho-syntaxique, de la détection des expressions multi-mots et de la lemmatisation pour un espace vectoriel continu. Pour l'évaluation, nous utilisons ces vecteurs sur une tâche de classification de phrases et les comparons avec le vecteur
du système de base. Nous explorons également l'approche d'adaptation de domaine pour construire des vecteurs. Malgré un petit nombre de vocabulaires et la petite taille du corpus d'apprentissage, les vecteurs spécialisés par domaine obtiennent de meilleures performances que les vecteurs hors domaine.

Les mots en arabe sont très proches lexicalement les uns des autres. La probabilité de tomber sur un mot correct en commettant une erreur typographique est plus importante que pour le français ou

pour l'anglais. Nous nous intéressons dans cet article à détecter les erreurs orthographiques plus précisément, celles générant des mots lexicalement corrects mais causant un dérèglement sémantique au niveau de la phrase. Nous décrivons et comparons deux méthodes se basant sur la représentation vectorielle du sens des mots. La première méthode utilise l'analyse sémantique latente (LSA). La seconde s'appuie sur le modèle Word2Vec et plus particulièrement l'architecture Skip-Gram. Les expérimentations ont montré que Skip-Gram surpasse LSA.

Cet article propose une approche d'analyse de sentiments à base d'aspects dans un texte d'opinion. Cette approche se base sur deux étapes principales : l'extraction d'aspects et la classification du sentiment relatif à chaque aspect. Pour l'extraction d'aspects, nous proposons une nouvelle approche
qui combine un CNN pour l'apprentissage de représentation de caractères, un b-LSTM pour joindre l'apprentissage de représentation de caractères et de mots et un CRF pour l'étiquetage des séquences
de mots en entités. Pour la classification de sentiments, nous utilisons un réseau à mémoire d'attention pour associer un sentiment (positif, négatif ou neutre) à une expression d'aspect donnée. Les
expérimentations sur des corpus d'avis (publics et industriels) en langue française ont montré des performances qui dépassent les méthodes existantes.

La Similarité Textuelle Sémantique (STS) est la base de nombreuses applications dans le Traitement
Automatique du Langage Naturel (TALN). Notre système combine des réseaux neuronaux convolutifs
et récurrents pour mesurer la similarité sémantique des phrases. Il utilise un réseau convolutif pour tenir compte du contexte local des mots et un LSTM pour prendre en considération le contexte

global d'une phrase. Cette combinaison des réseaux préserve mieux les informations significatives des phrases et améliore le calcul de la similarité entre les phrases. Notre modèle a obtenu de bons résultats et est compétitif avec les meilleurs systèmes de l'état de l'art.

Les méthodes d'évaluation actuelles des représentations vectorielles de mots utilisent généralement un jeu de données restreint et biaisé. Pour pallier à ce problème nous présentons une nouvelle approche, basée sur la similarité entre les synsets associés aux mots dans la volumineuse base de données lexicale WordNet. Notre méthode d'évaluation consiste dans un premier temps à classer automatiquement les représentions vectorielles de mots à l'aide d'un algorithme de clustering, puis à évaluer la cohérence sémantique et syntaxique des clusters produits. Cette évaluation est effectuée en calculant la similarité entre les mots de chaque cluster, pris deux à deux, en utilisant des mesures
de similarité entre les mots dans WordNet proposées par NLTK (wup _similarity). Nous obtenons, pour chaque cluster, une valeur entre 0 et 1. Un cluster dont la valeur est 1 est un cluster dont tous les mots appartiennent au même synset. Nous calculons ensuite la moyenne des mesures de tous les clusters. Nous avons utilisé notre nouvelle approche pour étudier et comparer trois méthodes de représentations vectorielles : une méthode traditionnelle, WebSOM et deux méthodes récentes, word2vec (Skip-Gram et CBOW) et GloVe, sur trois corpus : en anglais, en français et en arabe.

Les corpus annotés en sens sont des ressources cruciales pour la tâche de désambiguïsation lexicale
(Word Sense Disambiguation). La plupart des langues n'en possèdent pas ou trop peu pour pouvoir construire des systèmes robustes. Nous nous intéressons ici à la langue arabe et présentons 12 corpus
annotés en sens, fabriqués automatiquement à partir de 12 corpus en langue anglaise. Nous évaluons

la qualité de nos systèmes de désambiguïsation grâce à un corpus d'évaluation en arabe nouvellement
disponible.

Un mésusage apparaît lorsqu'un patient ne respecte pas sa prescription et fait des actions pouvant mener à des effets nocifs. Bien que ces situations soient dangereuses, les patients ne signalent généralement pas les mésusages à leurs médecins. Il est donc nécessaire d'étudier d'autres sources
d'information pour découvrir ce qui se passe en réalité. Nous proposons d'étudier les forums de santé
en ligne. L'objectif de notre travail consiste à explorer les forums de santé avec des méthodes de classification supervisée afin d'identifier les messages contenant un mésusage de médicament. Notre
méthode permet de détecter les mésusages avec une F-mesure allant jusqu'à 0,810. Cette méthode peut aider dans la détection de mésusages et la construction d'un corpus exploitable par les experts pour étudier les types de mésusages commis par les patients.

L'identification des entités nommées dans un texte est une étape fondamentale pour de nombreuses
tâches d'extraction d'information. Pour avoir une identification complète, une étape de désambiguïsation des entités similaires doit être réalisée. Celle-ci s'appuie souvent sur la seule description
textuelle des entités. Or, les bases de connaissances contiennent des informations plus riches, sous la
forme de relations entre les entités : cette information peut également être exploitée pour améliorer la désambiguïsation des entités. Nous proposons dans cet article une approche d'apprentissage de

représentations distribuées de ces relations et leur utilisation pour la tâche de désambiguïsation d'entités nommées. Nous montrons le gain de cette méthode sur un corpus d'évaluation standard, en
anglais, issu de la tâche de désambiguïsation d'entités de la campagne TAC-KBP.

Nous étudions la possibilité de construire un dispositif de traduction automatique neuronale du japonais vers le français, capable d'obtenir des résultats à la hauteur de l'état de l'art, sachant que l'on ne peut disposer de grands corpus alignés bilingues. Nous proposons un état de l'art et relevons
de nombreux signes d'amélioration de la qualité des traductions, en comparaison aux traductions statistiques jusque-là prédominantes. Nous testons ensuite un des baselines librement disponibles, OpenNMT, qui produit des résultats encourageants. Sur la base de cette expérience, nous proposons
plusieurs pistes pour améliorer à terme la traduction et pour compenser le manque de corpus.

Au-delà des modèles destinés à construire des plongements lexicaux à partir de corpus, des méthodes
de spécialisation de ces représentations selon différentes orientations ont été proposées. Une part importante d'entre elles repose sur l'utilisation de connaissances externes. Dans cet article, nous proposons Pseudofit, une nouvelle méthode de spécialisation de plongements lexicaux focalisée sur la similarité sémantique et opérant sans connaissances externes. Pseudofit s'appuie sur la notion de pseudo-sens afin d'obtenir plusieurs représentations pour un même mot et utilise cette pluralité pour rendre plus génériques les plongements initiaux. Nous illustrons l'intérêt de Pseudofit pour l'extraction de synonymes et nous explorons dans ce cadre différentes variantes visant à en améliorer
les résultats.

Pour synthétiser automatiquement et de manière expressive des livres audio, il est nécessaire de connaître le type des discours à oraliser. Ceci étant, dans un roman ou une nouvelle, les perspectives
narratives et les types de discours évoluent souvent entre de la narration, du récitatif, du discours direct, du discours rapporté, voire des dialogues. Dans ce travail, nous allons présenter un outil qui a été développé à partir de l'analyse d'un corpus de livres audio (extraits de Madame Bovary et
des Mystères de Paris) et qui prend comme unité de base pour l'analyse le paragraphe. Cet outil permet donc non seulement de déterminer automatiquement les types de discours (narration, discours
direct, dialogue), et donc de savoir qui parle, mais également d'annoter l'extension des modifications discursives. Ce dernier point est important, notamment dans le cas d'incises de citation où le narrateur
reprend la parole dans une séquence au discours direct. Dans sa forme actuelle, l'outil atteint un taux
de 89 % de bonne détection.

Ce travail présente une étude de l'impact du prétraitement linguistique (suppression de mots vides, racinisation et détection d'emoji, de négation et d'entités nommées) sur la classification des sentiments
en dialecte Tunisien. Nous évaluons cet impact sur trois corpus de tailles et contenus différents. Deux
techniques de classification sont utilisées : Naïve bayes et Support Vector Machines. Nous comparons
nos résultats aux résultats de référence obtenus sur ces même corpus. Nos résultats soulignent l'impact

positif de la phase de prétraitement sur la performance de la classification.

Dans cet article, nous proposons plusieurs approches pour l'identification automatique de phrases parallèles qui nécessitent du contexte linguistique extra-phrastique pour être correctement traduites. Notre objectif à long terme est de construire de façon automatique un jeu de test de phrases dépendantes du contexte afin d'évaluer les modèles de traduction automatique conçus pour améliorer la
traduction de phénomènes discursifs et contextuels. Nous fournissons une discussion et une critique
qui montrent que les approches actuelles ne nous permettent pas d'atteindre notre but et qui suggère
que l'évaluation individuelle de phenomènes est probablement la meilleure solution.

Dans une conversation humain-humain entre un usager et un interlocuteur en centre d'assistance, on
se place dans le contexte où l'issue du dialogue est caractérisée par une notion de succès ou d'échec,
explicitement annotée ou extrapolée. L'étude envisage différents paramètres susceptibles d'exercer une influence sur un modèle de classification prédictive des échecs constatés. On cherchera d'une part à exploiter une modélisation de la distribution lexicale tirant parti de l'asymétrie des rôles des locuteurs. On examinera d'autre part si la partie du lexique plus étroitement liée au domaine d'assistance client abordé ici, modifie la qualité de la prédiction. On interrogera enfin les perspectives de généralisation du modèle à des corpus morphologiquement comparables.

La détection automatique de la négation fait souvent partie des pré-requis dans les systèmes d'extraction d'information, notamment dans le domaine biomédical. Cet article présente nos

contributions
concernant la détection de la portée de la négation en français et portugais brésilien. Nous présentons
d'une part deux corpus principalement constitués d'extraits de protocoles d'essais cliniques en français et portugais brésilien, dédiés aux critères d'inclusion de patients. Les marqueurs de négation et
leurs portées y ont été annotés manuellement. Nous présentons d'autre part une approche par réseau
de neurones récurrents pour extraire les portées.

Le projet PASTEL étudie l'acceptabilité et l'utilisabilité des transcriptions automatiques dans le cadre d'enseignements magistraux. Il s'agit d'outiller les apprenants pour enrichir de manière synchrone et automatique les informations auxquelles ils peuvent avoir accès durant la séance. Cet enrichissement
s'appuie sur des traitements automatiques du langage naturel effectués sur les transcriptions automatiques. Nous présentons dans cet article un travail portant sur l'annotation d'enregistrements de
cours magistraux enregistrés dans le cadre du projet CominOpenCourseware. Ces annotations visent
à effectuer des expériences de transcription automatique, segmentation thématique, appariement
automatique en temps réel avec des ressources externes... Ce corpus comprend plus de neuf heures
de parole annotées. Nous présentons également des expériences préliminaires réalisées pour évaluer
l'adaptation automatique de notre système de reconnaissance de la parole.

Cet article a pour but de montrer la faisabilité d'un système de fouille de texte pour alimenter un moteur d'inférences capable de construire, à partir de prédicats extraits des articles scientifiques, un réseau de signalisation en biologie systémique. Cette fouille se réalise en deux étapes : la recherche
de phrases d'intérêt dans un grand corpus scientifique, puis la construction automatique de prédicats.
Ces deux étapes utilisent un système de cascades de transducteurs.

Nous présentons une méthode pour extraire des couples de termes médicaux translittérés de l'anglais
en caractères arabes. Nous avons proposé un processus de construction des translittérations de termes
anglais en arabe. Celui-ci s'appuie sur une étude en corpus pour la création d'une table de correspondances des caractères anglais en arabe mais aussi sur des règles de conversion qui tiennent compte de
certaines particularités de la langue arabe comme l'agglutination et la non-voyellation. Nous avons évalué l'apport de l'utilisation de la translittération pour identifier des couples de termes anglais
arabe sur un corpus parallèle de textes médicaux. Les résultats montrent que parmi 137 couples de mots anglais-arabe extraits, 120 sont jugés corrects (soit 87,59%), dont 107 représentent des couples
de termes médicaux (soit 89,16% des translittérations correctes et 78,10% des résultats).

Dans le domaine médical, la simplification des textes est à la fois une tâche souhaitable pour les patients et scientifiquement stimulante pour le domaine du traitement automatique du langage naturel.
En effet, les comptes rendus médicaux peuvent être difficile à comprendre pour les non spécialistes,

essentiellement à cause de termes médicaux spécifiques (prurit, par exemple). La substitution de ces
termes par des mots du langage courant peut aider le patient à une meilleure compréhension. Dans cet article, nous présentons une méthode de simplification dans le domaine médical (en français) basée sur un réseau lexico-sémantique. Nous traitons cette difficulté sémantique par le remplacement
du terme médical difficile par un synonyme ou terme qui lui est lié sémantiquement à l'aide d'un réseau lexico-sémantique français. Nous présentons dans ce papier, une telle méthode ainsi que son
évaluation.

L'estimation contrastive bruitée (NCE) et l'échantillonage par importance (IS) sont des procédures d'entraînement basées sur l'échantillonage, que l'on utilise habituellement à la place de l'estimation du maximum de vraisemblance (MLE) pour éviter le calcul du softmax lorsque l'on entraîne des modèles de langue neuronaux. Dans cet article, nous cherchons à résumer le fonctionnement de ces
algorithmes, et leur utilisation dans la littérature du TAL. Nous les comparons expérimentalement, et présentons des manières de faciliter l'entraînement du NCE.

Nous nous intéressons dans cet article à l'extraction de thèmes à partir de retranscriptions textuelles de réunions. Ce type de corpus est bruité, il manque de formatage, il est peu structuré avec plusieurs
locuteurs qui interviennent et l'information y est souvent éparpillée. Nous présentons une étude expérimentale utilisant des méthodes fondées sur la mesure tf-idf et l'extraction de topics sur un corpus réel de référence (le corpus AMI) pour l'étude de réunions. Nous comparons nos résultats avec les résumés fournis par le corpus.

L'alternance codique est le phénomène qui consiste à alterner les langues au cours d'une même conversation ou d'une même phrase. Avec l'augmentation du volume généré par les utilisateurs, ce phénomène essentiellement oral, se retrouve de plus en plus dans les textes écrits, nécessitant d'adapter les tâches et modèles de traitement automatique de la langue à ce nouveau type d'énoncés.
Ce travail présente la collecte et l'annotation en partie du discours d'un corpus d'énoncés comportant
des alternances codiques et évalue leur impact sur la tâche d'analyse morpho-syntaxique.

Dans cet article, nous comparons l'impact de la simplification d'un schéma d'annotation sur un système de repérage d'entités nommées (REN). Une simplification consiste à rassembler les types d'entités nommées (EN) sous deux types génériques (personne et lieu), l'autre revient à mieux définir
chaque type d'EN. Nous observons une amélioration des résultats sur les deux versions simplifiées. Nous étudions également la possibilité de retrouver le niveau de détail des types d'EN du schéma d'origine à partir des versions simplifiées. L'utilisation de règles de conversion permet de recouvrer les types d'EN d'origine, mais il reste une forme d'ambiguïté contextuelle qu'il est impossible de lever au moyen de règles.

Une simple détection d'opinions positives ou négatives ne satisfait plus les chercheurs et les entreprises.
Le monde des affaires est à la recherche d'un «aperçu des affaires». Beaucoup de méthodes peuvent
être utilisées pour traiter le problème. Cependant, leurs performances, lorsque les classes ne sont pas

équilibrées, peuvent être dégradées. Notre travail se concentre sur l'étude des techniques visant à traiter les données déséquilibrées en parfumerie. Cinq méthodes ont été comparées : Smote, Adasyn,
Tomek links, Smote-TL et la modification du poids des classe. L'algorithme d'apprentissage choisi est le SVM et l'évaluation est réalisée par le calcul des scores de précision, de rappel et de f-mesure.
Selon les résultats expérimentaux, la méthode en ajustant le poids sur des coût d'erreurs avec SVM,
nous permet d'obtenir notre meilleure F-mesure.

Détecter la complexité lexicale est une étape importante pour la simplification automatique de textes, servant lors de l'identification des éléments lexicaux à substituer. Dans ce travail, nous explorons l'utilité des plongements lexicaux pour mesurer la complexité de mots en français, en les combinant avec d'autres traits reconnus comme étant utiles pour cette tâche. Nos résultats sur une tâche d'ordonnancement de synonymes selon leur complexité montrent que les plongements seuls donnent de meilleurs résultats que nombreux autres traits, bien que leur performance reste inférieure
à celle de systèmes basés sur la fréquence pour cette langue.

Dans cet article, nous présentons une approche hybride pour la translitération de l'arabizi algérien. Nous avons élaboré un ensemble de règles permettant le passage de l'arabizi vers l'arabe. Á partir de ces règles nous générons un ensemble de candidats pour la translitération de chaque mot en arabizi vers l'arabe, et un parmi ces candidats sera ensuite identifié et extrait comme le meilleur candidat. Cette approche a été expérimentée en utilisant trois corpus de tests. Les résultats obtenus montrent une amélioration du score de précision qui était pour le meilleur des cas de l'ordre de 75,11%. Ces résultats ont aussi permis de vérifier que notre approche est très compétitive par

rapport aux travaux traitant de la translitération de l'arabizi en général.

Les lieux constituent une information structurante de nombreux textes (récits, romans, articles journalistiques, guides touristiques, itinéraires de randonnées, etc.) et leur recensement et leur analyse doit tenir compte des aspects thématiques abordés dans les textes. Le travail proposé ici s'inscrit dans les domaines de la linguistique de corpus et de la cartographie. La définition de lieu est augmentée de celle d'objet localisé et la désignation de ces lieux peut alors être construite sur un nom propre ou un nom commun. Des expérimentations sont menées afin d'identifier les lieux noms propres avec des gazetiers et les lieux noms communs grâce à un modèle d'apprentissage automatique. Les résultats sont discutés sous la forme d'une comparaison entre les caractéristiques linguistiques des noms de lieux et les propriétés visuelles que devront satisfaire leur représentation cartographique.

Les systèmes de désambigu??sation d'entités nommées utilisent principalement sur des bases de connaissances encyclopédique telles que DBpedia ou Freebase. Dans ce papier, nous utilisons à la place, un réseau léxicalo-sémantique nommé JeuxDeMots pour conjointement désambigu??ser et typer
les entités nommées. Notre approche combine les plongements de mots et la similitude de chemins dans un graphe résultant à des résultats encourageants sur un ensemble de documents provenant du
journal Le Monde.

L'utilisation des emojis dans les messageries sociales n'a eu de cesse d'augmenter ces dernières années. Plusieurs travaux récents ont porté sur la prédiction d'emojis afin d'épargner à l'utillisateur le
parcours de librairies d'emojis de plus en plus conséquentes. Nous proposons une méthode

permettant
de récupérer automatiquement les catégories d'emojis à partir de leur contexte d'utilisation afin d'améliorer la prédiction finale. Pour ce faire nous utilisons des plongements lexicaux en considérant
les emojis comme des mots présents dans des tweets. Nous appliquons ensuite un regroupement automatique restreint aux emojis visages afin de vérifier l'adéquation des résultats avec la théorie d'Ekman. L'approche est reproductible et applicable sur tous types d'emojis, ou lorsqu'il est nécessaire
de prédire de nombreuses classes.

Dans cet article, nous abordons le problème de la détection de la polarité pour l'analyse de sentiments
au niveau des aspects dans un contexte bilingue : nous proposons d'adapter le composant de détection
de polarité d'un système préexistant d'analyse de sentiments au niveau des aspects, très performant
pour la tâche, et reposant sur l'utilisation de ressources sémantiques riches pour une langue donnée, à
une langue sémantiquement moins richement dotée. L'idée sous-jacente est de réduire le besoin de supervision nécessaire à la construction des ressources sémantiques essentielles à notre système. À
cette fin, la langue source, peu dotée, est traduite vers la langue cible, et les traductions parallèles sont ensuite alignées mot à mot. Les informations sémantiques riches sont alors extraites de la langue
cible par le système de détection de polarité, et ces informations sont ensuite alignées vers la langue

source. Nous présentons les différentes étapes de cette expérience, ainsi que l'évaluation finale. Nous
concluons par quelques perspectives.

Dans cet article, nous étudions la contribution de propriétés syntaxiques à la tâche de clustering d'instances de relations sémantiques. Les instances, constituées de couples de concepts apparaissant dans des textes scientifiques, sont représentées dans une matrice où on les croise avec une représentation de leur contexte de co-occurrence. Différentes variantes de représentations sont envisagées pour ce contexte, en faisant appel à la fouille de données séquentielles et à l'analyse syntaxique en dépendances. Nos comparaisons suggèrent que les attributs issus d'analyses syntaxiques permettent d'améliorer la qualité du clustering final.

Ce travail montre que la dégradation des performances souvent observée lors de l'application d'un analyseur morpho-syntaxique à des données hors domaine résulte souvent d'incohérences entre les annotations des ensembles de test et d'apprentissage. Nous montrons comment le principe de variation des annotations, introduit par Dickinson & Meurers (2003) pour identifier automatiquement les erreurs d'annotation, peut être utilisé pour identifier ces incohérences et évaluer leur impact sur les performances des analyseurs morpho-syntaxiques.

Les conversations techniques en ligne sont un type de productions linguistiques qui par de nombreux
aspects se démarquent des objets plus usuellement étudiés en traitement automatique des langues : il
s'agit de dialogues écrits entre deux locuteurs qui servent de support à la résolution coopérative des problèmes des usagers. Nous proposons de décrire ici ces conversations par un étiquetage en actes de

dialogue spécifiquement conçu pour les conversations en ligne. Différents systèmes de prédictions ont été évalués ainsi qu'une méthode permettant de s'abstraire des spécificités lexicales du corpus d'apprentissage.

Suite à la mise en place d'une chaîne traitement destinée à extraire automatiquement des actions de
maintenance réalisées sur des composants dans des comptes rendus, nous avons cherché à constituer
des ressources lexicales à partir de textes souvent mal normalisés sur le plan linguistique. Nous avons ainsi développé une application web, CuriosiText, qui permet de lancer un traitement Word2Vec et de peupler semi automatiquement une ontologie métier avec les termes similaires correctement détectés. Des relations métiers spécifiques peuvent également être ajoutées.

La plateforme ACCOLÉ (Annotation Collaborative d'erreurs de traduction pour COrpus aLignÉs) propose une palette de services innovants permettant de répondre aux besoins modernes d'analyse d'erreurs de traduction : gestion simplifiée des corpus et des typologies d'erreurs, annotation d'erreurs efficace, collaboration et/ou supervision lors de l'annotation, recherche de modèle d'erreurs dans les annotations.

LIMA est un analyseur linguistique libre d'envergure industrielle. Nous présentons ici ses évolutions depuis la dernière publication en 2014.

Dans cet article, nous présentons un outil pour effectuer l'étiquetage rapide de textes bruts. Il peut charger des documents annotés depuis divers formats, notamment BRAT et GATE. Il se base sur

des

raccourcis claviers intuitifs et la diffusion d'annotation à l'échelle du document. Il permet d'entraîner des systèmes par apprentissage que l'on peut alors utiliser pour préannoter les textes.

OntoNotes comprend le seul corpus manuellement annoté en sens librement disponible pour l'arabe.

Elle reste peu connue et utilisée certainement parce que le projet s'est achevé sans lier cet inventaire

au Princeton WordNet qui lui aurait ouvert l'accès à son riche écosystème. Dans cet article, nous présentons une version étendue de OntoNotes Release 5.0 que nous avons créée en suivant une méthodologie de construction semi-automatique. Il s'agit d'une mise à jour de la partie arabe annotée

en sens du corpus en ajoutant l'alignement vers le Princeton WordNet 3.0. Cette ressource qui comprend plus de 12 500 mots annotés est librement disponible pour la communauté. Nous espérons

qu'elle deviendra un standard pour l'évaluation de la désambiguïsation lexicale de l'arabe.

Le domaine médical fait partie de la vie quotidienne pour des raisons de santé, mais la disponibilité des informations médicales ne garantit pas leur compréhension correcte par les patients. Plusieurs études ont démontré qu'il existe une difficulté réelle dans la compréhension de contenus médicaux par les patients. Nous proposons d'exploiter les méthodes d'oculométrie pour étudier ces questions et pour détecter quelles unités linguistiques posent des difficultés de compréhension. Pour cela, des textes médicaux en version originale et simplifiée sont exploités. L'oculométrie permet de suivre le regard des participants de l'étude et de révéler les indicateurs de lecture, comme la durée des fixations,

les régressions et les saccades. Les résultats indiquent qu'il existe une différence statistiquement

significative lors de la lecture des versions originales et simplifiées des documents de santé testés.

Nous proposons dans cet article une adaptation de l'approche compositionnelle étendue capable d'aligner des termes de longueurs variables à partir de corpus comparables, en modifiant la représentation des termes complexes. Nous proposons également de nouveaux modes de pondération pour
l'approche standard qui améliorent les résultats des approches état de l'art pour les termes simples et
complexes en domaine de spécialité.

Les modèles vectoriels de sémantique distributionnelle (ou word embeddings), notamment ceux produits par les méthodes neuronales, posent des questions de reproductibilité et donnent des représentations différentes à chaque utilisation, même sans modifier leurs paramètres. Nous présentons
ici un ensemble d'expérimentations permettant de mesurer cette instabilité, à la fois globalement et localement. Globalement, nous avons mesuré le taux de variation du voisinage des mots sur trois corpus différents, qui est estimé autour de 17% pour les 25 plus proches voisins d'un mot. Localement,
nous avons identifié et caractérisé certaines zones de l'espace sémantique qui montrent une relative
stabilité, ainsi que des cas de grande instabilité.

Les nourrissons doivent trouver des limites de mots dans le flux continu de la parole. De nombreuses
études computationnelles étudient de tels mécanismes. Cependant, la majorité d'entre elles se sont concentrées sur l'anglais, une langue morphologiquement simple et qui rend la tâche de

segmentation aisée. Les langues polysynthétiques - pour lesquelles chaque mot est composé de plusieurs
morphèmes - peuvent présenter des difficultés supplémentaires lors de la segmentation. De plus, le mot est considéré comme la cible de la segmentation, mais il est possible que les nourrissons segmentent des morphèmes et non pas des mots. Notre étude se concentre sur deux langues ayant des
structures morphologiques différentes, le chintang et le japonais. Trois algorithmes de segmentation conceptuellement variés sont évalués sur des représentations de mots et de morphèmes. L'évaluation
de ces algorithmes nous mène à tirer plusieurs conclusions. Le modèle lexical est le plus performant,
notamment lorsqu'on considère les morphèmes et non pas les mots. De plus, en faisant varier leur évaluation en fonction de la langue, le japonais nous apporte de meilleurs résultats.

Le nouvel état de l'art en traduction automatique (TA) s'appuie sur des méthodes neuronales, qui diffèrent profondément des méthodes utilisées antérieurement. Les métriques automatiques classiques
sont mal adaptées pour rendre compte de la nature du saut qualitatif observé. Cet article propose un
protocole d'évaluation pour la traduction de l'anglais vers le français spécifiquement focalisé sur la compétence morphologique des systèmes de TA, en étudiant leurs performances sur différents phénomènes grammaticaux.

Cet article présente une nouvelle méthode d'étiquetage en parties du discours adaptée aux langues peu dotées : la définition du contexte utilisé pour construire les plongements lexicaux est adaptée à la tâche, et de nouveaux vecteurs sont créés pour les mots inconnus. Les expériences menées sur

le

picard, le malgache et l'alsacien montrent que cette méthode améliore l'état de l'art pour ces trois langues peu dotées.

Dans cet article, nous présentons une modélisation de la situation d'acquisition de la syntaxe de sa langue maternelle par un enfant inspirée des "jeux de langages" de Luc Steels. Le modèle suppose que l'enfant a accès à une représentation sémantique des énoncés qui lui sont adressés, et qu'il doit réagir en désignant la tête syntaxique de ces énoncés. Nous décrivons des expériences exploitant des

données du corpus CHILDES et mettant en jeu un processus d'acquisition simple mais efficace.

Cet article présente un système open source et modulaire pour le résumé automatique : MOTS, développé en Java. Son architecture permet d'implémenter et tester de nouvelles méthodes de résumé

automatique et de les comparer avec des méthodes existantes dans un cadre unifié. Ce système, le premier complètement modulaire pour le résumé automatique permet à l'heure actuelle de définir plus

de cent combinaisons de modules afin de résumer automatiquement des textes en langage naturel.

Construire des systèmes de dialogue qui conversent avec les humains afin de les aider dans leurs tâches quotidiennes est devenu une priorité. Certains de ces systèmes produisent des dialogues en cherchant le meilleur énoncé (réponse) parmi un ensemble d'énoncés candidats. Le choix de la réponse est conditionné par l'historique de la conversation appelé contexte. Ces systèmes ordonnent

les énoncés candidats par leur adéquation au contexte, le meilleur est ensuite choisi. Les approches

existantes à base de réseaux de neurones profonds sont performantes pour cette tâche. Dans cet article,
nous améliorons une approche état de l'art à base d'un dual encodeur LSTM. En se basant sur la similarité sémantique entre le contexte et la réponse, notre approche apprend à mieux distinguer les bonnes réponses des mauvaises. Les résultats expérimentaux sur un large corpus de chats d'Ubuntu
montrent une amélioration significative de 7, 6 et 2 points sur le Rappel1, 2 et 5) respectivement
par rapport au meilleur système état de l'art.

Les approches neuronales obtiennent depuis plusieurs années des résultats intéressants en extraction d'événements. Cependant, les approches développées dans ce cadre se limitent généralement à
un contexte phrastique. Or, si certains types d'événements sont aisément identifiables à ce niveau, l'exploitation d'indices présents dans d'autres phrases est parfois nécessaire pour permettre de désambiguïser des événements. Dans cet article, nous proposons ainsi l'intégration d'une représentation
d'un contexte plus large pour améliorer l'apprentissage d'un réseau convolutif. Cette représentation est obtenue par amorçage en exploitant les résultats d'un premier modèle convolutif opérant au niveau
phrastique. Dans le cadre d'une évaluation réalisée sur les données de la campagne TAC 2017, nous
montrons que ce modèle global obtient un gain significatif par rapport au modèle local, ces deux modèles étant eux-mêmes compétitifs par rapport aux résultats de TAC 2017. Nous étudions également en détail le gain de performance de notre nouveau modèle au travers de plusieurs expériences
complémentaires.

Les registres de langue sont un trait stylistique marquant dans l'appréciation d'un texte ou d'un discours. Cependant, il sont encore peu étudiés en traitement automatique des langues. Dans cet article,
nous présentons une approche semi-supervisée permettant la construction conjointe d'un corpus de textes étiquetés en registres et d'un classifieur associé. Cette approche s'appuie sur un ensemble initial
et restreint de données expertes. Via une collecte automatique et massive de pages web, l'approche
procède par itérations en alternant l'apprentissage d'un classifieur intermédiaire et l'annotation de nouveaux textes pour augmenter le corpus étiqueté. Nous appliquons cette approche aux registres familier, courant et soutenu. À l'issue du processus de construction, le corpus étiqueté regroupe 800 000 textes et le classifieur, un réseau de neurones, présente un taux de bonne classification de 87 %.

En désambiguïsation lexicale, l'utilisation des réseaux de neurones est encore peu présente et très récente. Cette direction est pourtant très prometteuse, tant les résultats obtenus par ces premiers systèmes arrivent systématiquement en tête des campagnes d'évaluation, malgré une marge d'amélioration qui semble encore importante. Nous présentons dans cet article une nouvelle architecture
à base de réseaux de neurones pour la désambiguïsation lexicale. Notre système est à la fois moins complexe à entraîner que les systèmes neuronaux existants et il obtient des résultats état de l'art sur la
plupart des tâches d'évaluation de la désambiguïsation lexicale en anglais. L'accent est porté sur la reproductibilité de notre système et de nos résultats, par l'utilisation d'un modèle de vecteurs de mots,

de corpus d'apprentissage et d'évaluation librement accessibles.

La désambiguïsation des rattachements prépositionnels est une tâche syntaxique qui demande des connaissances sémantiques, pouvant être extraites d'une image associée au texte traité. Nous présentons et analysons les difficultés de cette tâche pour laquelle nous construisons un système complet
entraîné sur une version étendue des annotations du corpus Flickr30k Entities. Lorsque la sémantique
lexicale n'est pas disponible, l'information visuelle apporte 3 % d'amélioration.

L'absence de données annotées peut être une difficulté majeure lorsque l'on s'intéresse à l'analyse de
documents manuscrits anciens. Pour contourner cette difficulté, nous proposons de diviser le problème
en deux, afin de pouvoir s'appuyer sur des données plus facilement accessibles. Dans cet article
nous présentons la partie décodeur d'un encodeur-décodeur multimodal utilisant l'apprentissage
par transfert de connaissances pour la transcription des titres de pièces de la Comédie Italienne. Le décodeur transforme un vecteur de n-grammes au niveau caractères en une séquence de caractères
correspondant à un mot. L'apprentissage par transfert de connaissances est réalisé principalement à
partir d'une nouvelle ressource inexploitée contemporaine à la Comédie-Italienne et thématiquement
proche ; ainsi que d'autres ressources couvrant d'autres domaines, des langages différents et même
des périodes différentes. Nous obtenons 97,27% de caractères bien reconnus sur les données de la

Comédie-Italienne, ainsi que 86,57% de mots correctement générés malgré une couverture de 67,58%
uniquement entre la Comédie-Italienne et l'ensemble d'apprentissage. Les expériences montrent qu'un tel système peut être une approche efficace dans le cadre d'apprentissage par transfert.

Malgré les faiblesses connues de cette métrique, les performances de différents systèmes de reconnaissance automatique de la parole sont généralement comparées à l'aide du taux d'erreur sur les mots.
Les transcriptions automatiques de ces systèmes sont de plus en plus exploitables et utilisées dans des
systèmes complexes de traitement automatique du langage naturel, par exemple pour la traduction automatique, l'indexation, la recherche documentaire... Des études récentes ont proposé des métriques
permettant de comparer la qualité des transcriptions automatiques de différents systèmes en fonction
de la tâche visée. Dans cette étude nous souhaitons mesurer, qualitativement, l'apport de l'adaptation
automatique des modèles de langage au domaine visé par un cours magistral. Les transcriptions du discours de l'enseignant peuvent servir de support à la navigation dans le document vidéo du cours magistral ou permettre l'enrichissement de son contenu pédagogique. C'est à-travers le prisme de ces
deux tâches que nous évaluons l'apport de l'adaptation du modèle de langage. Les expériences ont été menées sur un corpus de cours magistraux et montrent combien le taux d'erreur sur les mots est
une métrique insuffisante qui masque les apports effectifs de l'adaptation des modèles de langage.

Nous nous intéressons ici à l'analyse de conversation par chat dans un contexte orienté-tâche avec un conseiller technique s'adressant à un client, où l'objectif est d'étiqueter les énoncés en actes de dialogue, pour alimenter des analyses des conversations en aval. Nous proposons une méthode légèrement supervisée à partir d'heuristiques simples, de quelques annotations de développement, et une méthode d'ensemble sur ces règles qui sert à annoter automatiquement un corpus plus large de façon bruitée qui peut servir d'entrainement à un modèle supervisé. Nous comparons cette approche à une approche supervisée classique et montrons qu'elle atteint des résultats très proches, à un coût moindre et tout en étant plus facile à adapter à de nouvelles données.

L'avènement des techniques d'apprentissage automatique profond a fait naître un besoin énorme de données d'entraînement. De telles données d'entraînement sont extrêmement coûteuses à créer, surtout lorsqu'une expertise dans le domaine est requise. L'une de ces tâches est l'apprentissage de la structure sémantique du discours, tâche très complexe avec des structures récursives avec des données éparses, mais qui est essentielle pour extraire des informations sémantiques profondes du texte. Nous décrivons nos expérimentations sur l'attachement des unités discursives pour former une structure, en utilisant le paradigme du data programming dans lequel peu ou pas d'annotations sont utilisées pour construire un ensemble de données d'entraînement "bruité". Le corpus de dialogues utilisé illustre des

contraintes à la fois linguistiques et non-linguistiques intéressantes qui doivent être apprises. Nous nous concentrons sur la structure des règles utilisées pour construire un modèle génératif et montrons
la compétitivité de notre approche par rapport à l'apprentissage supervisé classique.

La compréhension automatique de texte est une tâche faisant partie de la famille des systèmes de Question/Réponse où les questions ne sont pas à portée générale mais sont liées à un document
particulier. Récemment de très grand corpus (SQuAD, MS MARCO) contenant des triplets (document,
question, réponse) ont été mis à la disposition de la communauté scientifique afin de développer des
méthodes supervisées à base de réseaux de neurones profonds en obtenant des résultats prometteurs.
Ces méthodes sont cependant très gourmandes en données d'apprentissage, données qui n'existent
pour le moment que pour la langue anglaise. Le but de cette étude est de permettre le développement
de telles ressources pour d'autres langues à moindre coût en proposant une méthode générant de manière semi-automatique des questions à partir d'une analyse sémantique d'un grand corpus. La collecte de questions naturelle est réduite à un ensemble de validation/test. L'application de cette méthode sur le corpus CALOR-Frame a permis de développer la ressource CALOR-QUEST présentée
dans cet article.

Le travail décrit le développement d'un chunker pour l'oral par apprentissage supervisé avec les

CRFs, à partir d'un corpus de référence de petite taille et composé de productions de nature différente : monologue préparé vs discussion spontanée. La méthodologie respecte les spécificités des données traitées. L'apprentissage tient compte des résultats proposés par différents étiqueteurs morpho-syntaxiques disponibles sans correction manuelle de leurs résultats. Les expériences montrent que le genre de discours (monologue vs discussion), la nature de discours (spontané vs préparé) et la taille du corpus peuvent influencer les résultats de l'apprentissage, ce qui confirme que la nature des données traitées est à prendre en considération dans l'interprétation des résultats.

En vue de distinguer la traduction littérale des autres procédés de traduction, des traducteurs et linguistes ont proposé plusieurs typologies pour caractériser les différents procédés de traduction, tels que l'équivalence idiomatique, la généralisation, la particularisation, la modulation sémantique, etc. En revanche, les techniques d'extraction de paraphrases à partir de corpus parallèles bilingues n'ont pas exploité ces informations. Dans ce travail, nous proposons une classification automatique des procédés de traduction en nous basant sur des exemples annotés manuellement dans un corpus parallèle (anglais-français) de TED Talks. Même si le jeu de données est petit, les résultats expérimentaux sont encourageants, et les expériences montrent la direction à suivre dans les futurs travaux.

L'objectif de ce travail est de présenter plusieurs observations, sur l'évaluation des analyseurs morpho-syntaxique en français, visant à remettre en cause le cadre habituel de l'apprentissage statistique

dans lequel les ensembles de test et d'apprentissage sont fixés arbitrairement et indépendemment du
modèle considéré. Nous montrons qu'il est possible de considérer des ensembles de test plus petits que ceux généralement utilisés sans conséquences sur la qualité de l'évaluation. Les exemples ainsi
« économisés » peuvent être utilisés en apprentissage pour améliorer les performances des systèmes
notamment dans des tâches d'adaptation au domaine.

Les interactions aliments-médicaments (FDI) se produisent lorsque des aliments et des médicaments
sont pris simultanément et provoquent un effet inattendu. Nous considérons l'extraction de ces interactions dans les textes comme une tâche d'extraction de relation pouvant être résolue par des méthodes de classification. Toutefois, étant donné que ces interactions sont décrites de manière très fine, nous sommes confrontés au manque de données et au manque d'exemples par type de relation. Pour résoudre ce problème, nous proposons d'appliquer une adaptation de domaine à partir des interactions médicament-médicament (DDI) qui est une tâche similaire, afin d'établir une correspondance entre les types de relations et d'étiqueter les instances FDI selon les types DDI. Notre
approche confirme une cohérence entre les 2 domaines et fournit une base pour la spécification des relations et la pré-annotation de nouvelles données. Les performances des modèles de classification
appuie également l'efficacité de l'adaptation de domaine sur notre tâche.

Les méthodes de recherche d'information permettent d'explorer les données textuelles. Nous les exploitons pour la détection de messages avec la non-adhérence médicamenteuse dans les forums

de discussion. La non-adhérence médicamenteuse correspond aux cas lorsqu'un patient ne respecte
pas les indications de son médecin et modifie les prises de médicaments (augmente ou diminue les doses, par exemple). Le moteur de recherche exploité montre 0,9 de précision sur les 10 premiers résultats avec un corpus équilibré, et 0,4 avec un corpus respectant la distribution naturelle des messages, qui est très déséquilibrée en défaveur de la catégorie recherchée. La précision diminue avec l'augmentation du nombre de résultats considérés alors que le rappel augmente. Nous exploitons
également le moteur de recherche sur de nouvelles données et avec des types précis de non-adhérence.

Les phrases parallèles contiennent des informations identiques ou très proches sémantiquement et offrent des indications importantes sur le fonctionnement de la langue. Lorsque les phrases sont différenciées par leur registre (comme expert vs. non-expert), elles peuvent être exploitées pour la simplification automatique de textes. Le but de la simplification automatique est d'améliorer la compréhension de textes. Par exemple, dans le domaine biomédical, la simplification peut permettre aux patients de mieux comprendre les textes relatifs à leur santé. Il existe cependant très peu de ressources pour la simplification en français. Nous proposons donc d'exploiter des corpus comparables, différenciés par leur technicité, pour y détecter des phrases parallèles et les aligner. Les
données de référence sont créées manuellement et montrent un accord inter-annotateur de 0,76. Nous
expérimentons sur des données équilibrées et déséquilibrées. La F-mesure sur les données équilibrées
atteint jusqu'à 0,94. Sur les données déséquilibrées, les résultats sont plus faibles (jusqu'à 0,92 de F-mesure) mais restent compétitifs lorsque les modèles sont entraînés sur les données équilibrées.

Nous décrivons dans cet article notre travail de développement d'un lexique morphologique et syntaxique à grande échelle de l'ancien français pour le traitement automatique des langues. Nous nous sommes appuyés sur des ressources dictionnairiques et lexicales dans lesquelles l'extraction d'informations structurées et exploitables a nécessité des développements spécifiques. De plus, la mise en correspondance d'informations provenant de ces différentes sources a soulevé des difficultés.
Nous donnons quelques indications quantitatives sur le lexique obtenu, et discutons de sa fiabilité dans sa version actuelle et des perspectives d'amélioration permises par l'existence d'une première version, notamment au travers de l'analyse automatique de données textuelles.

L'apprentissage par transfert est une solution au problème de l'apprentissage de systèmes de traduction
automatique neuronaux pour des paires de langues peu dotées. Dans cet article, nous proposons une
analyse de cette méthode. Nous souhaitons évaluer l'impact de la quantité de données et celui de la proximité des langues impliquées pour obtenir le meilleur transfert possible. Nous prenons en compte
ces deux paramètres non seulement pour une tâche de traduction "classique" mais également lorsque les corpus de données font défaut. Enfin, il s'agit de proposer une approche où volume de données et proximité des langues sont combinées afin de ne plus avoir à trancher entre ces deux éléments.

L'évaluation de plongements issus de réseaux de neurones est un procédé complexe. La qualité des
plongements est liée à la tâche spécifique pour laquelle ils ont été entraînés et l'évaluation de cette

tâche peut être un procédé long et onéreux s'il y a besoin d'annotateurs humains. Il peut donc être préférable d'estimer leur qualité grâce à des mesures objectives rapides et reproductibles sur des tâches annexes. Cet article propose une méthode générique pour estimer la qualité d'un plongement.
Appliquée à la synthèse de parole par sélection d'unités guidée par réseaux de neurones, cette méthode
permet de comparer deux systèmes distincts.

L'apprentissage par transfert représente la capacité qu'un modèle neuronal entraîné sur une tâche à généraliser suffisamment et correctement pour produire des résultats pertinents sur une autre tâche proche mais différente. Nous présentons dans cet article une approche fondée sur l'apprentissage par transfert pour construire automatiquement des outils d'analyse de textes des réseaux sociaux en exploitant les similarités entre les textes d'une langue bien dotée (forme standard d'une langue) et les textes d'une langue peu dotée (langue utilisée en réseaux sociaux). Nous avons expérimenté notre approche sur plusieurs langues ainsi que sur trois tâches d'annotation linguistique (étiquetage morpho-syntaxique, annotation en parties du discours et reconnaissance d'entités nommées). Les résultats obtenus sont très satisfaisants et montrent l'intérêt de l'apprentissage par transfert pour tirer profit des modèles neuronaux profonds sans la contrainte d'avoir à disposition une quantité de données importante nécessaire pour avoir une performance acceptable.

Nous présentons ici une exploration préliminaire du concept d'informativité ?la quantité d'information qu'une phrase fournit sur l'un des mots qui le compose? et ses usages potentiels pour l'apprentissage
de plongements de mots robustes à partir de données en faible quantité. Une mesure d'informativité est prédite à partir d'algorithmes de classification de phrases, que nous comparons à une série de phrases annotées manuellement. Nous concluons que ces deux mesures correspondent à des

notions

différentes d'informativité. Néanmoins, nos expériences montrent que la prédiction extraite de la classification a un impact sur la qualité des plongements de mots lors de l'apprentissage.

Dans le contexte médical, un patient ou médecin virtuel dialoguant permet de former les apprenants au diagnostic médical via la simulation de manière autonome. Dans ce travail, nous avons exploité les propriétés sémantiques capturées par les représentations distribuées de mots pour la recherche de questions similaires dans le système de dialogues d'un agent conversationnel médical. Deux systèmes de dialogues ont été créés et évalués sur des jeux de données collectées lors des tests avec les apprenants. Le premier système fondé sur la correspondance de règles de dialogue créées la main présente une performance globale de 92% comme taux de réponses cohérentes sur le cas clinique étudié tandis que le second système qui combine les règles de dialogue et la similarité sémantique réalise une performance de 97% de réponses cohérentes en réduisant de 7% les erreurs de compréhension par rapport au système de correspondance de règles.

Les méthodes endogènes se trouvent au coeur de la construction des ressources de connaissance telles

que les réseaux lexico-sémantiques. Dans le cadre de l'expérience décrite dans le présent article, nous nous focalisons sur les méthodes d'inférence des relations. Nous considérons, en particulier, les cas d'inférence des relations sémantiques et des raffinements de sens. Les différents mécanismes

d'inférence des relations sémantiques y compris dans le contexte de polysémie de termes ont été décrits par Zarrouk (2015) pour le contexte monolingue. À notre connaissance, il n'existe pas

de travaux concernant l'inférence des relations sémantiques et des raffinements dans le contexte d'amélioration d'une ressource multilingue.

Cet article présente la constitution d'un corpus de textes produits, sur des données lors de dictées, par des enfants paralysés cérébraux (PC) ou dysorthographiques, son annotation en termes d'erreurs
orthographiques, et enfin son analyse quantitative. Cette analyse de corpus a pour objectif de définir des besoins réels en matière de correction orthographique, et ce pour les personnes souffrant de troubles du langage écrit comme pour le grand public. Notre étude suggère que les correcteurs orthographiques ne répondent que partiellement à ces besoins.

Les schémas Winograd sont des problèmes de résolution d'anaphores conçus pour nécessiter un raisonnement sur des connaissances du monde. Par construction, ils sont insensibles à des statistiques
simples (co-occurrences en corpus). Pourtant, aujourd'hui, les systèmes état de l'art pour l'anglais se
basent sur des modèles de langue pour résoudre les schémas (Trinh & Le, 2018). Nous présentons dans cet article une étude visant à tester des modèles similaires sur les schémas en français. Cela nous conduit à revenir sur les métriques d'évaluation utilisées dans la communauté pour les schémas
Winograd. Les performances que nous obtenons, surtout comparées à celles de Amsili & Seminck (2017b), suggèrent que l'approche par modèle de langue des schémas Winograd reste limitée, sans doute en partie à cause du fait que les modèles de langue encodent très difficilement le genre de raisonnement nécessaire à la résolution des schémas Winograd.

Ces dernières années, les recherches sur la fouille d'opinions ou l'analyse des sentiments sont menées
activement dans le domaine du Traitement Automatique des Langues (TAL). De nombreuses études

scientifiques portent sur l'extraction automatique des opinions positives ou négatives et de leurs cibles. Ce travail propose d'identifier automatiquement une évaluation, exprimée explicitement ou implicitement par des internautes dans le corpus d'avis tiré du Web. Six catégories d'évaluation sont proposées : opinion positive, opinion négative, opinion mixte, intention, suggestion et description. La méthode utilisée est fondée sur l'apprentissage supervisé qui tient compte des caractéristiques linguistiques de chaque catégorie retenue. L'une des difficultés que nous avons rencontrée concerne
le déséquilibre entre les classes d'évaluation créées, cependant, cet obstacle a pu être surmonté dans
l'apprentissage grâce aux stratégies de sur-échantillonnage et aux stratégies algorithmiques.

Nos recherches sur la langue corse nous amènent naturellement à envisager l'utilisation d'outils pour
le traitement automatique du langage. Après une brève introduction sur le corse et sur le projet qui constitue notre cadre de travail, nous proposons un état des lieux concernant l'application du TAL aux
langues peu dotées, dont le corse. Nous définissons ensuite les actions qui peuvent être entreprises,
ainsi que la manière dont elles peuvent s'intégrer dans le cadre de notre projet, afin de progresser vers
la constitution de ressources et la construction d'outils pour le TAL corse.

La résolution d'anaphores est une tâche fondamentale pour la plupart des applications du TALN. Cette tâche reste un problème difficile qui nécessite plusieurs sources de connaissances et des techniques d'apprentissage efficaces, notamment pour la langue arabe. Cet article présente une nouvelle approche de résolution d'anaphores pronominales dans les textes arabes en se basant sur

une méthode d'Apprentissage par Renforcement AR qui utilise l'algorithme Q-learning. Le processus de résolution comporte une étape d'identification des pronoms et des antécédents candidats et une autre de résolution. L'algorithme Q-learning permet d'apprendre dans un environnement dynamique et incertain. Il cherche à optimiser pour chaque pronom anaphorique, une séquence de choix de critères pour évaluer les antécédents et sélectionner le meilleur. Le système de résolution est évalué sur des textes littéraires, des textes journalistiques et des manuels techniques. Le taux de précision atteint jusqu'à 77,14%.

Les mesures de similarité textuelle ont une place importante en TAL, du fait de leurs nombreuses applications, en recherche d'information et en classification notamment. En revanche, le dialogue fait moins l'objet d'attention sur cette question. Nous nous intéressons ici à la production d'une similarité dans le contexte d'un corpus de conversations par chat à l'aide de méthodes non-supervisées,
exploitant à différents niveaux la notion de sémantique distributionnelle, sous forme d'embeddings. Dans un même temps, pour enrichir la mesure, et permettre une meilleure interprétation des résultats,
nous établissons des alignements explicites des tours de parole dans les conversations, en exploitant
la distance de Wasserstein, qui permet de prendre en compte leur dimension structurelle. Enfin, nous évaluons notre approche à l'aide d'une tâche externe sur la petite partie annotée du corpus, et observons qu'elle donne de meilleurs résultats qu'une variante plus naïve à base de moyennes.

L'avènement des approches neuronales de bout en bout a entraîné une rupture dans la façon dont était
jusqu'à présent envisagée et implémentée la tâche de résolution des coréférences. Nous pensons que

cette rupture impose de remettre en question la conception des mentions en termes de syntagmes maximaux, au moins pour certaines applications dont nous donnons deux exemples. Dans cette perspective, nous proposons une nouvelle formulation de la tâche, basée sur les têtes, accompagnée
d'une adaptation du modèle de Lee et al. (2017) qui l'implémente.

Nous nous intéressons dans cet article à la problématique de réutilisation de textes dans les livres liturgiques du Moyen Âge. Plus particulièrement, nous étudions les variations textuelles de la prière Obsecro Te souvent présente dans les livres d'heures. L'observation manuelle de 772 copies de l'Obsecro Te a montré l'existence de plus de 21 000 variantes textuelles. Dans le but de pouvoir les extraire automatiquement et les catégoriser, nous proposons dans un premier temps une classification
lexico-sémantique au niveau n-grammes de mots pour ensuite rendre compte des performances de plusieurs approches état-de-l'art d'appariement automatique de variantes textuelles de l'Obsecro Te.

Cet article présente un retour d'expérience sur la transformation de corpus annotés pour l'alsacien et
l'occitan vers le format CONLL-U défini dans le projet Universal Dependencies. Il met en particulier l'accent sur divers points de vigilance à prendre en compte, concernant la tokénisation et la définition
des catégories pour l'annotation.

Les corpus annotés sont des ressources difficiles à créer en raison du grand effort humain qu'elles impliquent. Une fois rendues disponibles, elles sont difficilement modifiables et tendent à ne pas évoluer pas dans le temps. Dans cet article, nous présentons un corpus annoté pour la

reconnaissance
des entités nommées libre et évolutif en utilisant les textes d'articles Wikinews français de 2016 à 2018, pour un total de 1191 articles annotés. Nous décrivons succinctement le guide d'annotation avant de situer notre corpus par rapport à d'autres corpus déjà existants. Nous donnerons également
un accord intra-annotateur afin de donner un indice de stabilité des annotations ainsi que le processus
global pour poursuivre les travaux d'enrichissement du corpus.

Cet article 1 propose une approche hybride pour la segmentation de documents basée sur l'agrégation
de différentes solutions. Divers algorithmes de segmentation peuvent être utilisés dans le système, ce qui permet la combinaison de stratégies multiples (spécifiques au domaine, supervisées et non-supervisées). Un ensemble de documents étiquetés, segmentés au préalable et représentatif du domaine
ciblé, doit être fourni pour être utilisé comme ensemble d'entraînement pour l'apprentissage des méthodes supervisées, et aussi comme ensemble de test pour l'évaluation de la performance de chaque
méthode, ce qui déterminera leur poids lors de la phase d'agrégation. L'approche proposée présente
de bonnes performances dans un scénario expérimental issu d'un corpus extrait du domaine juridique.

L'usage, le sens et la connotation des mots peuvent changer au cours du temps. Les plongements lexicaux diachroniques permettent de modéliser ces changements de manière non supervisée. Dans

cet article nous étudions l'impact de plusieurs fonctions de coût sur l'apprentissage de plongements dynamiques, en comparant les comportements de variantes du modèle Dynamic Bernoulli Embeddings.

Les plongements dynamiques sont estimés sur deux corpus couvrant les mêmes deux décennies, le New York Times Annotated Corpus en anglais et une sélection d'articles du journal Le Monde en français, ce qui nous permet de mettre en place un processus d'analyse bilingue de l'évolution de l'usage des mots.

La littérature des réseaux complexes a montré la pertinence de l'étude de la langue sous forme de réseau pour différentes applications : désambiguïsation, résumé automatique, classification des langues, etc. Cette même littérature a démontré que les réseaux de co-occurrences de mots possèdent une structure de communautés latente. Nous formulons l'hypothèse que cette structuration du réseau sous forme de communautés est utile pour travailler sur la sémantique d'une langue et introduisons donc dans cet article une méthode d'apprentissage de plongements originale basée sur cette hypothèse.

Cette hypothèse est cohérente avec la proximité qui existe entre la détection de communautés sur un réseau de co-occurrences et la factorisation d'une matrice de co-occurrences, méthode couramment utilisée pour l'apprentissage de plongements lexicaux. Nous décrivons notre méthode structurée en trois étapes : construction et pré-traitement du réseau, détection de la structure de communautés, construction des plongements de mots à partir de cette structure. Après avoir décrit cette nouvelle méthodologie, nous montrons la pertinence de notre approche avec des premiers résultats d'évaluation

sur les tâches de catégorisation et de similarité. Enfin, nous discutons des perspectives importantes d'un tel modèle issu des réseaux complexes : les dimensions du modèle (les communautés) semblent
interprétables, l'apprentissage est rapide, la construction d'un nouveau plongement est presque instantanée, et il est envisageable d'en expérimenter une version incrémentale pour travailler sur des
corpus textuels temporels.

Nous présentons une étude visant à comparer 11 différents analyseurs en dépendances du français sur un corpus spécialisé (constitué des archives des articles de la conférence TALN). En l'absence de gold standard, nous utilisons chacune des sorties de ces analyseurs pour construire des thésaurus
distributionnels en utilisant une méthode à base de fréquence. Nous comparons ces 11 thésaurus afin
de proposer un premier aperçu de l'impact du choix d'un analyseur par rapport à un autre.

En Désambiguïsation Lexicale (DL), les systèmes supervisés dominent largement les campagnes d'évaluation. La performance et la couverture de ces systèmes sont cependant rapidement limités par la faible quantité de corpus annotés en sens disponibles. Dans cet article, nous présentons deux
nouvelles méthodes qui visent à résoudre ce problème en exploitant les relations sémantiques entre les
sens tels que la synonymie, l'hyperonymie et l'hyponymie, afin de compresser le vocabulaire de sens
de WordNet, et ainsi réduire le nombre d'étiquettes différentes nécessaires pour pouvoir désambiguïser

tous les mots de la base lexicale. Nos méthodes permettent de réduire considérablement la taille des
modèles de DL neuronaux, avec l'avantage d'améliorer leur couverture sans données supplémentaires,
et sans impacter leur précision. En plus de nos méthodes, nous présentons un système de DL qui tire
parti des récents travaux sur les représentations vectorielles de mots contextualisées, afin d'obtenir des résultats qui surpassent largement l'état de l'art sur toutes les tâches d'évaluation de la DL.

Les corpus textuels sont utiles pour diverses applications de traitement automatique des langues (TAL) en fournissant les données nécessaires pour leur création, adaptation ou évaluation. Cependant,
dans certains domaines comme le domaine médical, l'accès aux données est rendu compliqué, voire
impossible, pour des raisons de confidentialité et d'éthique. Il existe néanmoins de réels besoins
en corpus cliniques pour l'enseignement et la recherche. Pour répondre à ce défi, nous présentons dans cet article le corpus CAS contenant des cas cliniques de patients, réels ou fictifs, que nous avons compilés. Ces cas cliniques en français couvrent plusieurs spécialités médicales et focalisent donc sur différentes situations cliniques. Actuellement, le corpus contient 4 300 cas (environ 1,5M d'occurrences de mots). Il est accompagné d'informations (discussions des cas cliniques, mots-clés, etc.) et d'annotations que nous avons effectuées au regard des besoins de la recherche en TAL dans ce domaine. Nous présentons également les résultats de premières expériences de recherche et
d'extraction d'information qui ont été effectuées avec ce corpus annoté. Ces expériences peuvent fournir une baseline à d'autres chercheurs souhaitant travailler avec les données.

Dans cet article, nous présentons une approche de bout en bout d'extraction de concepts sémantiques
de la parole. En particulier, nous mettons en avant l'apport d'une chaîne d'apprentissage successif pilotée par une stratégie de curriculum d'apprentissage. Dans la chaîne d'apprentissage mise en place,
nous exploitons des données françaises annotées en entités nommées que nous supposons être des
concepts plus génériques que les concepts sémantiques liés à une application informatique spécifique.
Dans cette étude, il s'agit d'extraire des concepts sémantiques dans le cadre de la tâche MEDIA.
Pour renforcer le système proposé, nous exploitons aussi des stratégies d'augmentation de données,
un modèle de langage 5-gramme, ainsi qu'un mode étoile aidant le système à se concentrer sur les concepts et leurs valeurs lors de l'apprentissage. Les résultats montrent un intérêt à l'utilisation des données d'entités nommées, permettant un gain relatif allant jusqu'à 6,5 %.

Cet article présente une méthodologie de détection des ellipses en anglais qui repose sur des patrons
combinant des informations sur les tokens, leur étiquette morphosyntaxique et leur lemme. Les
patrons sont évalués sur deux corpus de sous-titres. Ces travaux constituent une étape préalable à une
étude contrastive et multi-genres de l'ellipse.

La génération automatique de poésie est une tâche ardue pour un système informatique. Pour qu'un
poème ait du sens, il est important de prendre en compte à la fois des aspects linguistiques et

littéraires.

Ces dernières années, un certain nombre d'approches fructueuses sont apparues, capables de modéliser
de manière adéquate divers aspects du langage naturel. En particulier, les modèles de langue basés
sur les réseaux de neurones ont amélioré l'état de l'art par rapport à la modélisation prédictive de langage, tandis que les topic models sont capables de capturer une certaine cohérence thématique. Dans cet article, on explorera comment ces approches peuvent être adaptées et combinées afin de modéliser les aspects linguistiques et littéraires nécessaires pour la génération de poésie. Le système
est exclusivement entraîné sur des textes génériques, et sa sortie est contrainte afin de conférer un caractère poétique au vers généré. Le cadre présenté est appliqué à la génération de poèmes en français, et évalué à l'aide d'une évaluation humaine.

Nous proposons une architecture neuronale avec les caractéristiques principales des modèles neuronaux de ces dernières années : les réseaux neuronaux récurrents bidirectionnels, les modèles encodeur-décodeur, et le modèle Transformer. Nous évaluons nos modèles sur trois tâches d'étiquetage de séquence, avec des résultats aux environs de l'état de l'art et souvent meilleurs, montrant ainsi l'intérêt
de cette architecture hybride pour ce type de tâches.

Nous présentons la base PolylexFLE, contenant 4295 expressions polylexicales. Elle est integrée dans
une plateforme d'apprentissage du FLE, SimpleApprenant, destinée à l'apprentissage des expressions
polylexicales verbales (idiomatiques, collocations ou expressions figées). Afin de proposer des

exercices adaptés au niveau du Cadre européen de référence pour les langues (CECR), nous avons
utilisé une procédure mixte (manuelle et automatique) pour annoter 1098 expressions selon les niveaux de compétence du CECR. L'article se concentre sur la procédure automatique qui identifie, dans un premier temps, les expressions de la base PolylexFLE dans un corpus à l'aide d'un système
à base d'expressions régulières. Dans un second temps, leur distribution au sein de corpus, annoté selon l'échelle du CECR, est estimée et transformée en un niveau CECR unique.