Résumés de thèses

Rubrique préparée par Fiammetta Namer

Université de Lorraine, UMR « ATILF »
Fiammetta.Namer@univ-lorraine.fr

Clémentine ADAM : (adam@univ-tlse2.fr)

Titre : Voisinage lexical pour l'analyse du discours

Mots-clés : analyse du discours, cohésion lexicale, analyse distributionnelle.

Title: *Distributional neighborhood for discourse analysis*

Keywords: *discourse analysis, lexical cohesion, distributional analysis*.

Thèse de doctorat en Sciences du Langage, CLLE-ERSS UMR 5263, Université de Toulouse 2 – Le Mirail, Toulouse, sous la direction de Cécile Fabre (Pr, Université Toulouse 2), de Nicholas Asher (DR, IRIT-CNRS) et de Philippe Muller (MC, Université de Toulouse 3). Thèse soutenue le 28/09/2012.

Jury : Mme Cécile Fabre (Pr, Université Toulouse 2, codirectrice), M. Nicholas Asher (DR, IRIT-CNRS, codirecteur), M. Philippe Muller (MC, Université Toulouse 3, codirecteur), Mme Pascale Sébillot (Pr, INSA de Rennes/IRISA, présidente et rapporteur), M. Thierry Poibeau (DR, LaTTiCe-CNRS, rapporteur), M. Olivier Ferret (CR, CEA LIST/LVIC, examinateur).

Résumé **:** *Cette thèse s'intéresse au rôle de la cohésion lexicale dans différentes approches de l'analyse du discours. Nous y explorons deux hypothèses principales :*
– l'analyse distributionnelle, qui permet de rapprocher des unités lexicales sur la base des contextes syntaxiques qu'elles partagent, met au jour des relations sémantiques variées pouvant être exploitées pour la détection de la cohésion lexicale des textes ;
– les indices lexicaux constituent des éléments de signalisation de l'organisation du discours pouvant être exploités aussi bien à un niveau local (identification de relations rhétoriques entre constituants élémentaires du discours) qu'à un niveau global (repérage ou caractérisation de segments de niveau supérieur dotés d'une

fonction rhétorique et garantissant la cohérence et la lisibilité du texte, par exemple des passages ayant une unité thématique).
Concernant le premier point, nous montrons la pertinence d'une ressource distributionnelle pour l'appréhension d'une large gamme de relations impliquées dans la cohésion lexicale des textes. Nous présentons les méthodes de projection et de filtrage que nous avons mises en œuvre pour la production de sorties exploitables.

Concernant le second point, nous fournissons une série d'éclairages qui montrent l'apport d'une prise en compte réfléchie de la cohésion lexicale pour une grande variété de problématiques liées à l'étude et au repérage automatique de l'organisation textuelle : segmentation thématique de textes, caractérisation des structures énumératives, étude de la corrélation entre lexique et structure rhétorique du discours, et enfin détection de réalisations d'une relation de discours en particulier, la relation d'élaboration.

URL où la thèse pourra être téléchargée : http://clementine.adam.free.fr

---

François-Régis CHAUMARTIN: (frc@proxem.com)

Titre : Antelope, une plate-forme de TAL permettant d'extraire les sens du texte - Théorie et applications de l'interface syntaxe-sémantique

Mots-clés : traitement de corpus, analyse syntaxique et sémantique, lexique sémantique, acquisition de connaissances à large échelle, désambiguïsation, extraction d'information, résolution d'anaphores et de coréférences, Théorie Sens-Texte.

Title: *Antelope, a NLP platform for extracting meaning from text: theory and applications of the syntax-semantics interface*

Keywords: *corpora analysis, parsing, semantic role labeling, semantic lexicon, large-scale knowledge acquisition, disambiguation, information extraction, anaphora and coreference resolution, Meaning-Text Theory.*

Thèse de doctorat en linguistique théorique, descriptive et automatique, ALPAGE, UFR de Linguistique Informatique, Université de Paris 7, Paris, sous la direction de Sylvain Kahane (Pr, Université Paris 10). Thèse soutenue le 25/09/2012.

Jury : M. Sylvain Kahane (Pr, Université Paris 10, directeur), Mme Laurence Danlos (Pr, Université Paris 7, présidente), Mme Adeline Nazarenko (Pr, Université Paris 13, rapporteur), M Pierre Zweigenbaum (DR, LIMSI-CNRS, rapporteur),

M. Christian Jacquelinet (Médecin des hôpitaux, Agence de la biomédecine, Lim&Bio, examinateur), M. Guy Perrier (Pr, Université de Lorraine, examinateur).

Résumé **:** *Créer rapidement un analyseur sémantique dédié à une tâche particulière n'est pas une tâche aisée. En effet, composants d'analyse et ressources linguistiques sont souvent définis avec des formats incompatibles entre eux, ce qui en rend l'assemblage complexe. Nous souhaitons apporter une réponse opérationnelle à ce problème avec la plate-forme de traitement linguistique Antelope, dont cette thèse décrit les principes de conception et de réalisation. En partie basée sur la Théorie Sens-Texte (TST), Antelope permet une analyse syntaxique et sémantique de corpus de volume important ; son objectif est de « rendre calculable » du texte tout-venant : avis de consommateurs, textes encyclopédiques, documents RH, articles de presse… Pour cela, Antelope intègre :*
(i) plusieurs composants prêts à l'emploi, couvrant les principales tâches de TAL, qui interagissent au sein d'un modèle de données linguistiques unifié ;
(ii) un lexique sémantique multilingue à large couverture constitué à partir de différentes sources.
Un effort d'intégration permet d'offrir une plate-forme robuste et homogène ; l'ensemble constitue une interface syntaxico-sémantique opérationnelle. La thèse présente la plate-forme et la compare à d'autres projets de référence ; elle souligne les bonnes pratiques à prendre en termes d'architecture logicielle pour qu'un tel ensemble complexe reste maintenable ; elle introduit aussi une démarche semi-supervisée d'acquisition de connaissances à large échelle.

URL où la thèse pourra être téléchargée :
http://www.proxem.com/download/private/TheseFRC.pdf

Aurore KOEHL : (akoehl@atilf.fr)

Titre : La construction morphologique des noms désadjectivaux suffixés en français

Mots-clés : morphologie constructionnelle, règle de construction de lexèmes, noms désadjectivaux, noms de propriété, morphologie basée sur l'usage, données extensives

Title: *Suffixed adjective-based nouns in French.*

Keywords: *constructional morphology, word formation rule, deadjectival nouns, property nouns, usage-based morphology, extensive data.*

Thèse de doctorat en Sciences du Langage, UMR 7118 ATILF, école doctorale LTS, UFR Sciences du Langage, Université de Lorraine, Nancy, sous la direction de Fiammetta Namer (Pr, Université de Lorraine). Thèse soutenue le 30/11/2012.

Jury : Mme Fiammetta Namer, (Pr, Université de Lorraine, directrice), Mme Anne Carlier (Pr, Université de Lille 3, présidente), Mme Georgette Dal (Pr, Université de Lille 3, rapporteur), M. Nabil Hathout (DR, CLLE-ERSS, CNRS, rapporteur), Mme Marie-Laurence Knittel (MC-HDR, Université de Lorraine, examinatrice).

Résumé : *Menée dans le cadre de la morphologie lexématique, cette thèse a vocation de faire progresser la réflexion sur l'une des questions centrales de la morphologie constructionnelle des langues, à savoir les critères d'identification des règles de construction de lexèmes (RCL), en prenant le cas des noms désadjectivaux du français comme support. Nous traitons les suffixes -ité (banalité), -eur (blancheur), -esse (tendresse), -itude (amplitude), -ise (gourmandise) et -erie (niaiserie) qui sont utilisés comme exposants de règle dans la construction de noms sur la base de lexèmes adjectivaux. Les noms étudiés proviennent du* Trésor de la langue française *informatisé, du journal électronique* Le Monde *et de la Toile.*

Comment détermine-t-on les RCL ? Une première hypothèse est qu'à un exposant formel identifié correspond une RCL à laquelle s'oppose une seconde hypothèse selon laquelle à une seule RCL correspondent plusieurs exposants. Il s'agit de déterminer quelle est l'influence de la valeur des exposants dans le dénombrement des RCL. Cela implique (i) d'étudier les conditions de sélection des bases et (ii) d'étudier les critères aboutissant aux différentes formes de noms désadjectivaux. La première question relève d'une logique liée aux conditions d'application des règles, la seconde relève des motivations du locuteur/scripteur intervenant dans les conditions de concurrence entre les suffixes. Pour chaque suffixation, nous menons une étude sur la disponibilité de chaque suffixe, en comparant les noms contenus dans le Trésor de la langue française *et les créations des locuteurs/scripteurs en recourant au corpus électronique du journal* Le Monde *et à la Toile. Nous étudions également si les RCL subissent d'autres influences que celles des exposants, en analysant les contextes d'apparition des doublons de noms construits sur une même base adjectivale (e.g. tendresse et tendreté).*

Parallèlement à cette étude, nous avons créé une base de données morphologique des dérivations d'adjectif à nom (nommée MORDAN) qui enregistre 3 983 couples (adjectif, nom) assortis d'informations formelles, sémantiques, historiques et pragmatiques. Chaque nouvelle forme est accompagnée d'un contexte d'apparition qui permet son interprétation. Cette base de données est une ressource libre disponible à l'adresse https://arcas.atilf.fr/mordan/ et sous forme d'un tableur excel à https://sites.google.com/site/koehlaurore/.

URL où la thèse pourra être téléchargée :
https://sites.google.com/site/koehlaurore/

Émeline LECUIT : (emeline.lecuit@univ-tours.fr)

Titre : Les tribulations d'un nom propre en traduction. Étude contrastive du nom propre et de sa traduction à partir d'un corpus aligné de dix langues européennes

Mots-clés : noms propres, traduction, corpus multilingue parallèle aligné, annotation, alignement.

Title: *Tribulations of a Proper Name in Translation. A Contrastive Study of Proper Name and its Translation Based on an Aligned Corpus in ten European Languages*.

Keywords: *proper names, translation, aligned parallel multilingual corpus, annotation, alignment.*

Thèse de doctorat en Sciences du Langage, Laboratoire Ligérien de Linguistique, UMR CNRS 7270, département de Sciences du Langage, UFR Lettres et Langues, Université François-Rabelais, Tours, sous la direction de Denis Maurel (Pr, Université François-Rabelais) et de Duško Vitas (Pr, Université de Belgrade). Thèse soutenue le 30/11/2012.

Jury : M. Denis Maurel (Pr, Université François-Rabelais, codirecteur), M. Duško Vitas (Pr, Université de Belgrade, codirecteur), M. Jean-Louis Fournier (Pr, Université François-Rabelais, président), M. Christopher Gledhill (Pr, Université Paris-Diderot, rapporteur), M. Pavel Bradzil (Pr, Université de Porto, rapporteur), M. Geoffrey Williams (Pr, Université Bretagne-Sud, rapporteur), M. Denis Jamet (Pr, Université Jean-Moulin Lyon 3, examinateur).

Résumé **:** *Les noms propres sont omniprésents et intéressent, depuis des siècles, philosophes et linguistes. Le travail réalisé ici est une étude contrastive des noms propres en traduction, divisée en quatre parties.*
Les deux premières parties sont théoriques. La première partie traite de la notion de nom propre en linguistique anglaise et en linguistique française. La deuxième partie présente les différents procédés de traduction, illustrés par des exemples sur les noms propres.
Les deux parties suivantes sont expérimentales. La troisième partie détaille les différentes étapes de la constitution de notre corpus multilingue parallèle aligné et annoté, composé de onze versions du roman de Jules Verne, Le Tour du monde en quatre-vingts jours, *en dix langues européennes. La quatrième partie expose les résultats obtenus suite à l'observation du comportement des noms propres en traduction.*

Cette étude contredit souvent l'hypothèse largement répandue de leur intraduisibilité.

URL où la thèse pourra être téléchargée : s'adresser à l'auteur.

Damien NOUVEL : (damien.nouvel@inria.fr)

Titre : Reconnaissance des entités nommées par exploration de règles d'annotation

Mots-clés : traitement automatique des langues, fouille de données, entités nommées, règles d'annotation.

Title: *Named Entity Recognition by mining Annotation Rules.*

Keywords: *Natural Language Processing, Named Entities, Data Mining, Annotation Rules.*

Thèse de doctorat en Informatique, Laboratoire d'informatique, département d'informatique, UFR de Sciences et Techniques, Université François Rabelais, Tours, sous la direction de Jean-Yves Antoine (Pr, Université François Rabelais). Thèse soutenue le 20/11/2012.

Jury : M. Jean-Yves Antoine (Pr, Université François Rabelais, directeur), M. Bruno Crémilleux (Pr, Université de Caen, rapporteur), Mme Sophie Rosset (DR, LIMSI-CNRS, rapporteur), M. Frédéric Béchet (Pr, Université Aix-Marseille, examinateur), Mme Nathalie Friburger (MC, Université François Rabelais, examinatrice), M. Arnaud Soulet (MC, Université François Rabelais, examinateur).

Résumé : *Ces dernières décennies, le développement considérable des technologies de l'information et de la communication a modifié en profondeur la manière dont nous avons accès aux connaissances. Face à l'afflux de données et à leur diversité, il est nécessaire de mettre au point des technologies performantes et robustes pour y rechercher des informations. Les entités nommées (personnes, lieux, organisations, dates, expressions numériques, marques, fonctions, etc.) sont sollicitées afin de catégoriser, indexer ou, plus généralement, manipuler des contenus. Notre travail porte sur la reconnaissance et l'annotation de ces objets au sein de transcriptions d'émissions radiodiffusées ou télévisuelles. À cet effet, nous interprétons l'annotation des entités nommées comme une structuration locale. Nous pouvons alors nous appuyer sur les données pour extraire empiriquement des règles qui régissent l'apparition de marqueurs (ou balises) d'annotation.*

En première partie, nous abordons la problématique du traitement automatique des entités nommées. Nous présentons le cadre théorique et décrivons les analyses généralement conduites pour traiter le langage naturel. Nous discutons ensuite de la problématique des entités nommées (rétrospective des notions couvertes, typologies, évaluation et annotation) et proposons une caractérisation de leur nature linguistique. Nous concluons cette partie par un positionnement au regard des approches état de l'art et par notre proposition, centrée sur les marqueurs d'annotation. En deuxième partie, nous exposons le formalisme d'exploration de données que nous adoptons. Nous commençons par le situer au sein des méthodes de fouille de textes. Puis nous nous dotons d'un cadre formel pour explorer les motifs qui sont corrélés à un ou plusieurs marqueurs d'annotation, appelés « règles d'annotation ». Enfin, nous décrivons quelques modèles adéquats lorsqu'il s'agit d'utiliser ces règles pour annoter un texte donné. La dernière partie décrit le système implémenté (mXS) et les résultats obtenus. Nous détaillons en premier lieu les modules de traitement, ressources lexicales et corpus à notre disposition. Nous présentons ensuite la mise en œuvre et les résultats de l'extraction des règles d'annotation à partir des données. Enfin, nous fournissons des résultats chiffrés relatifs aux performances obtenues par mXS sur les données Ester2 et Étape, ainsi que des indicateurs supplémentaires quant au comportement du système à divers points de vue et dans diverses configurations. Ils montrent que l'approche que nous proposons est compétitive et qu'elle ouvre des perspectives dans le cadre du traitement du langage et de l'annotation automatique.

URL où la thèse pourra être téléchargée : http://damien.nouvels.net/fr/recherche

---

Ludovic TANGUY : (tanguy@univ-tlse2.fr)

Titre : Complexification des données et des techniques en linguistique : contributions du TAL aux solutions et aux problèmes

Mots-clés : TAL, linguistique outillée, corpus, annotation, analyses statistiques, apprentissage.

Title: *Complexity growth in linguistic data and techniques: NLP as a solution and a problem.*

Keywords: *NLP, corpus linguistics, data annotation, statistics, machine learning.*

Mémoire de HDR en Linguistique, UMR CLLE-ERSS, département des Sciences du Langage, UFR Langues, Lettres et Civilisations Étrangères, Toulouse, sous la direction d'Anne Condamines (DR, CLLE-ERSS). HDR soutenue le 11/09/2012.

Jury : Mme Anne Condamines (DR, CLLE-ERSS, directrice), M. Benoît Habert (Pr, ENS de Lyon, président), M. Mathieu Valette (Pr, INaLCO, rapporteur), M. François Yvon (Pr, Université de Paris-Sud, rapporteur), Mme Marie-Paule Péry-Woodley (Pr émérite, Université de Toulouse 2, examinatrice), Mme Pascale Sébillot (Pr, INSA de Rennes, examinatrice).

Résumé : *Ce mémoire d'habilitation est l'occasion de faire le bilan de mon activité d'enseignant-chercheur en traitement automatique des langues (TAL) dans un laboratoire de linguistique (CLLE-ERSS) et des principales évolutions de l'outillage informatique de la linguistique au cours des quinze dernières années.*

Mes recherches portent notamment sur le repérage de structures morphosyntaxiques dans les textes, l'analyse des structures du discours et l'acquisition de ressources lexicales à partir de corpus. Certaines se positionnent dans des cadres applicatifs comme la recherche d'information et la classification de textes, mais aussi dans des contextes plus spécifiques en lien avec d'autres disciplines (médecine, psychologie, sociologie...).

En m'appuyant sur la diversité de ces travaux et de mes collaborations, j'identifie quatre dimensions d'évolution principales :

- *l'augmentation de la masse de données langagières disponibles et notamment la part croissante de l'utilisation du Web comme corpus ;*
- *la complexification de l'outillage informatique disponible pour gérer la masse et la variété des données accessibles (outils de constitution et d'interrogation de corpus) ;*
- *la complexification de l'annotation des données langagières, qu'elle soit manuelle, assistée ou automatique ;*
- *la montée en puissance, en TAL mais aussi en linguistique descriptive, des méthodes quantitatives (depuis l'analyse statistique jusqu'aux techniques de fouille de données et d'apprentissage).*

Si les avancées techniques du TAL ont permis d'accroître de façon conséquente les potentialités d'investigation du matériau langagier, et dans certains cas de dégager de nouveaux questionnements, elles ont aussi contribué à creuser un fossé entre les deux composantes (informatique et linguistique) de la discipline.

À travers ma propre expérience d'acteur ou d'accompagnateur de ces changements et avec une vocation de « passeur » interdisciplinaire, je cherche à dégager les principaux enjeux actuels pour la linguistique outillée :

- *doter la linguistique descriptive d'outils de visualisation de données pour aborder la complexité, en exploitant les avancées théoriques et techniques de ce nouveau champ disciplinaire et en les adaptant aux spécificités du matériau langagier ;*
- *rendre abordables aux linguistes les techniques fondamentales de l'analyse statistique, mais aussi les méthodes d'apprentissage artificiel seules capables d'assister l'investigation et l'exploitation de données massives et complexes ;*

– *replacer la linguistique au sein des développements actuels du TAL, notamment par le biais de l'utilisation de descripteurs linguistiques riches dans les outils de traitement par apprentissage, pour un bénéfice mutuel.*

URL où l'HDR pourra être téléchargée : http://w3.erss.univ-tlse2.fr/membre/tanguy/HDR.html

---

Charles TEISSÈDRE : (charles.teissedre@gmail.com)

Titre : Analyse sémantique automatique des adverbiaux de localisation temporelle : application à la recherche d'information et à l'acquisition de connaissances

Mots-clés : extraction d'informations temporelles, annotation sémantique des adverbiaux de localisation temporelle, recherche d'information, acquisition de connaissances.

Title: *Automatic Semantic Analysis of Temporal Locating Adverbials: Application to Information Retrieval and Knowledge Acquisition.*

Keywords: *Temporal Information Extraction; Semantic Annotation of Temporal Locating Adverbials; Information Retrieval; Knowledge Acquisition.*

Thèse de doctorat en Sciences du Langage, spécialité traitement automatique des langues, UMR 7114 MoDyCo, département de Sciences du langage, UFR PHILLIA, Université Paris-Ouest Nanterre La Défense, Nanterre, sous la direction de Jean-Luc Minel (Pr, Université Paris-Ouest) et de Delphine Batistelli (MC-HDR, Université Paris-Sorbonne). Thèse soutenue le 22/11/2012.

Jury : M. Jean-Luc Minel (Pr, Université Paris-Ouest, co-directeur), Mme Delphine Batistelli (MC-HDR, Université Paris-Sorbonne, codirectrice), Mme Adeline Nazarenko (Pr, Université Paris-Nord, rapporteur et présidente), Mme Nathalie Aussenac-Gilles (DR, IRIT-CNRS, rapporteur), M. Guy Lapalme (Pr, Université de Montréal, examinateur), M. Maarten de Rijke (Pr, Université d'Amsterdam, examinateur), Mme Florence Amardeilh (DR, MONDECA, examinatrice).

Résumé : *Cette thèse aborde la question de l'accès aux textes numériques, en particulier de l'accès à leur « contenu informationnel », vu sous l'angle de l'ancrage temporel. Conciliant une approche linguistique et une approche applicative, ces travaux visent à contribuer à l'élaboration de nouveaux outils pour la fouille de textes, la recherche d'information et la gestion des connaissances – nouveaux outils en mesure de tirer parti de la sémantique des informations relatives au repérage temporel exprimées dans les textes. Il s'agit ainsi à la fois de*

mettre en œuvre des systèmes d'interaction avec les utilisateurs et de parvenir à modéliser la sémantique des unités textuelles qui contribuent de façon saillante à l'ancrage dans le temps des situations décrites dans les textes : les adverbiaux de localisation temporelle.

La représentation formelle que l'on en propose, qui procède d'une analyse linguistique, les décrit sous la forme d'une succession d'opérations sémantiques agissant sur un repère temporel noyau. Ces opérations visent à décrire la façon dont se détermine en langue la localisation temporelle à partir d'un repère initial, qui peut être formé d'une référence calendaire (« dès les années 20 »), mais également d'une référence déictique (« jusqu'à la semaine dernière »), anaphorique (« ce jour-là ») ou relative à un procès (« depuis la fin de la campagne électorale »). Notre proposition de modélisation permet de pouvoir décrire des informations présentant un certain degré d'indétermination sans en fermer l'interprétation (« jusque vers la fin des années 30 »). Elle permet également d'exprimer des informations en intension (« de février à août, tous les jours sauf le dimanche, de 10 h à 19 h »). Elle est ainsi plus expressive que les modèles généralement utilisés en ingénierie des connaissances – modèles qui, le plus souvent, représentent un repère temporel sous la forme d'une date ou d'un intervalle de dates.

Articulant notre proposition de modélisation avec les modèles standard des intervalles de dates, nous montrons qu'il devient possible d'élaborer des systèmes de recherche d'information susceptibles de traiter des requêtes associant un critère calendaire avec un ensemble de mots-clés, telles que « les universités au début du XII^e^ siècle » ou « le vote des femmes depuis 1900 », par exemple. On présente ainsi une heuristique permettant d'associer un intervalle calendaire à un adverbial de localisation temporelle, ainsi qu'un modèle de pertinence permettant de comparer des intervalles entre eux et de déterminer une mesure de similarité. Cette mesure permet d'ordonner par pertinence un ensemble de réponses pour une requête exprimant un critère calendaire.

S'appuyant sur les outils développés pour automatiser l'analyse (schéma d'annotation, ressources pour l'analyse sémantique, mesure de pertinence), on montre qu'il devient également possible d'interagir avec des données structurées décrivant des informations temporelles, à la fois pour les interroger et pour les enrichir de façon semi-automatique, afin, par exemple, de constituer des bases de connaissances.

URL où la thèse pourra être téléchargée : http://tel.archives-ouvertes.fr/tel-00762440

Juliette THUILIER : (juliette.thuilier@paris-sorbonne.fr)

Titre : Contraintes préférentielles et ordre des mots en français

Mots-clés : préférences syntaxiques, position de l'adjectif épithète, ordre des compléments verbaux, données de corpus, analyse de données, annotation sémantique, régression logistique, modèle à effets mixtes, questionnaire d'acceptabilité.

Title: *Soft constraints and word order in French.*

Keywords: *syntactic preferences, position of the attributive adjective, verbal complement ordering, corpus data, data analysis, semantic annotation, logistic regression, mixed-effects modelling, questionnaires of acceptability.*

Thèse de doctorat en Linguistique théorique, descriptive et automatique, UMRi001 ALPAGE (INRIA – Paris-Diderot), UFR de Linguistique, Université Paris-Diderot, Paris, sous la direction de Laurence Danlos (Pr, Université Paris-Diderot) et de Benoît Crabbé (MC, Université Paris-Diderot). Thèse soutenue le 28/09/2012.

Jury : Mme Laurence Danlos (Pr, Université Paris-Diderot, codirectrice), M. Benoît Crabbé (MC, Université Paris-Diderot, codirecteur), Mme Pollet Samvelian (Pr, Université Paris 3 Sorbonne Nouvelle, présidente), M. Philippe Blache (DR, LPL-CNRS, rapporteur), M. Shravan Vasishth (Pr, Université de Potsdam, rapporteur), Mme Anne Abeillé (Pr, Université Paris-Diderot, examinatrice).

Résumé **:** *Cette thèse propose une approche empirique et expérimentale de la syntaxe à travers l'étude de la notion de contrainte préférentielle (*soft constraint*) et son application à deux phénomènes d'ordre en français : la position de l'adjectif épithète ainsi que l'ordre relatif des deux compléments sous-catégorisés par le verbe et apparaissant en position postverbale. L'objectif est de proposer une étude détaillée de ces phénomènes linguistiques en s'appuyant sur une méthodologie particulière qui mêle l'analyse de données de corpus et les méthodes expérimentales.*

Le chapitre 1 est consacré aux contraintes préférentielles. Ces dernières sont conçues comme des contraintes qui n'affectent pas la grammaticalité mais l'acceptabilité des phrases. En nous appuyant sur une série de phénomènes observables dans différentes langues ou variétés de langues, nous émettons l'hypothèse selon laquelle ces contraintes constituent des propriétés spécifiques à la langue dont il faut rendre compte dans le champ de la syntaxe. Sur le plan méthodologique, l'étude de ces contraintes est rendue problématique par leur nature : étant donné qu'elles n'agissent pas sur la grammaticalité des phrases, les

contraintes préférentielles échappent aux méthodes traditionnelles de la syntaxe (introspection et jugement de grammaticalité). Il est donc nécessaire de définir des outils permettant leur description et leur analyse. C'est l'objet du chapitre 2. Les méthodes envisagées sont l'analyse statistique de données de corpus et, dans une moindre mesure, l'expérimentation psycholinguistique. Les outils dédiés à l'analyse de données de corpus, à savoir la régression logistique et les modèles à effets mixtes, sont présentés de façon détaillée dans ce chapitre. Nous exposons également les principes méthodologiques à l'œuvre dans le recueil et l'analyse des jugements de locuteurs collectés à l'aide de questionnaires.

En ce qui concerne la position de l'adjectif, le chapitre 3 reprend une grande partie des facteurs ayant été identifiés comme pouvant avoir une influence sur ce phénomène. Dans le chapitre 4, nous testons la plupart de ces contraintes et nous proposons une analyse statistique de données extraites du corpus French Treebank. *Nous montrons notamment l'importance de l'item adjectival ainsi que de l'item nominal avec lequel il se combine. Certaines contraintes syntaxiques concernant la configuration du syntagme adjectival et du syntagme nominal jouent également un rôle dans le choix de la position. L'ensemble de ces conclusions est formalisé sous la forme d'un modèle statistique qui permet d'estimer l'importance relative de chaque facteur. Les résultats de ce modèle sont mis en parallèle avec les préférences de 141 locuteurs recueillies à l'aide d'un questionnaire. La bonne corrélation entre préférences des locuteurs pour l'antéposition et la probabilité d'antéposition prédite par le modèle, montre que le modèle construit sur les données de corpus semble être en correspondance avec une forme de savoir langagier.*

Les deux derniers chapitres de la thèse sont consacrés au problème de l'ordre relatif des compléments du verbe. Ce phénomène, peu étudié pour le français, est présenté dans le chapitre 5. Nous exposons notamment le rôle de facteurs liés aux hiérarchies de poids, aux hiérarchies lexico-sémantiques et aux hiérarchies relatives au discours dans des phénomènes de linéarisation des dépendants du verbe dans d'autres langues que le français (notamment l'anglais et l'allemand).

*Le travail sur données attestées, présenté dans le chapitre 6, est mené sur un échantillon de phrases extraites de deux corpus journalistiques (*French Treebank *et* L'Est Républicain*) et de deux corpus de données orales (ESTER et C-ORAL-ROM). Nous montrons l'importante influence du poids des constituants dans le choix de l'ordre : dans une langue SVO comme le français, les séquences courtes précèdent les séquences longues dans plus de 86 % des cas. Nous mettons également en lumière le rôle fondamental du lemme verbal associé à sa classe sémantique, annotée à partir du dictionnaire de Dubois & Dubois-Charlier. Enfin, l'analyse des données de corpus ainsi que deux questionnaires d'élicitation de jugement d'acceptabilité ne permettent pas de montrer une influence du caractère animé, ni de l'opposition « donné/nouveau » sur l'ordre des compléments, contrairement à ce*

qui est observé dans d'autres langues comme l'anglais ou l'allemand. Le rôle du poids et des lemmes verbaux est capté grâce à une modélisation statistique qui permet notamment de déterminer l'ordre pour lequel chaque lemme associé à sa classe sémantique a une préférence, une fois les effets de poids pris en compte.

URL où la thèse pourra être téléchargée : http://www.linguist.univ-paris-diderot.fr/~jthuilier/