Résumés de thèses

Rubrique préparée par Fiammetta Namer

Université Nancy2 de Nancy, UMR « ATILF »
Fiammetta.Namer@univ-nancy2.fr

André BITTAR : (andre.bittar@linguist.jussieu.fr)

Titre : Construction d'un TimeBank du français : un corpus de référence annoté selon la norme ISO-TimeML.

Mots-clés : annotation temporelle, corpus annoté, temporalité, ISO-TimeML, événement, expression temporelle, relation temporelle, traitement automatique des langues.

Title : *Building a TimeBank for French : A Reference Corpus Annotated According to the ISO-TimeML Standard.*

Keywords : *temporal annotation, annotated corpus, time, ISO-TimeML, event, temporal expression, temporal relations, natural language processing.*

Thèse de doctorat en Linguistique théorique, descriptive et automatique, UFRL et UMRI Alpage-INRIA, Université de Paris 7, Parissous la codirection de Laurence Danlos (Pr, Alpage-INRIA & Université de Paris 7), Pascal Amsili (MC, LLF & Université Paris 7) et Pascal Denis (CR, Alpage-INRIA). Thèse soutenue le 19/11/2010.

Jury : Mme Laurence Danlos (Pr, Université de Paris 7, codirectrice), M. Pascal Amsili (MC, Université de Paris 7, codirecteur), M. Pascal Denis (CR, Alpage-INRIA, codirecteur), M. Philippe Müller (MC, IRIT & Université Paul-Sabatier, président), M. Michel Gagnon (Pr, École polytechnique de Montréal, rapporteur), M. James Pustejovsky (Pr, Brandeis University, rapporteur).

Résumé *: Cette thèse présente le développement de ressources pour le traitement des informations temporelles de textes en français et en particulier la construction d'un corpus de référence, le French TimeBank, annoté selon la norme ISO-TimeML. Pour la création de ce corpus, nous avons adopté une méthodologie qui adhère à des principes de bonne pratique bien établis dans le domaine de la linguistique de corpus. Les expressions temporelles, les événements ainsi que les relations temporelles qui existent entre ces entités y sont marqués. Pour la mise en*

œuvre de ce projet d'annotation, nous avons aussi développé un guide d'annotation ISO-TimeML pour le français et un système d'annotation automatique à base de règles. Ce système a servi à effectuer une préannotation automatique des textes du corpus avant une étape d'annotation manuelle. Les performances du système sont mesurées dans une évaluation.

Le guide d'annotation a été nécessaire pour diriger le processus d'annotation manuelle ainsi que pour caractériser les annotations du corpus. La création de ce guide a donné lieu à des améliorations du langage ISO-TimeML. D'une part, nous proposons des extensions du schéma d'annotation ISO-TimeML afin de permettre le traitement de phénomènes linguistiques en français, tels que les temps et modes verbaux, l'aspect grammatical et les verbes modaux. D'autre part, nous proposons un ensemble d'améliorations pouvant également s'appliquer à d'autres langues. Le système d'annotation automatique a quant à lui servi à effectuer un traitement préalable des textes avant de procéder à une correction par des annotateurs humains.

Une analyse quantitative et qualitative du French TimeBank nous a permis d'évaluer la méthodologie suivie pour sa création, y compris une évaluation des effets de la préannotation automatique. Cette analyse a également donné un aperçu du matériel linguistique employé pour l'expression de la temporalité en français et comprend une comparaison avec le corpus TimeBank 1.2 développé pour l'anglais. Nous relevons aussi un certain nombre de points pouvant aider à améliorer les outils d'annotations traditionnels.

URL où la thèse pourra être téléchargée : http://www.linguist.univ-paris-diderot.fr/~abittar/docs/Bittar-PhD.pdf

---

Frederik CAILLIAU : (cailliau@sinequa.com)

Titre : Des ressources aux traitements linguistiques : le rôle d'une architecture linguistique.

Mots-clés : architecture linguistique, ressource linguistique, gestion de ressources linguistiques, système de TAL, traitement automatique des langues.

Title : *The Role of a Linguistic Architecture in Language Processing and its Resources.*

Keywords : *linguistic architecture, linguistic resource, linguistic resource management, NLP system, NLP tool, natural language processing.*

Thèse de doctorat en Informatique, Université de Paris 13 – Paris Nord, Institut Galilée, LIPN – UMR CNRS 7030, Villetaneuse, sous la direction de Adeline Nazarenko (Pr, Université de Paris 13). Thèse soutenue le 09/12/2010.

Jury : Mme Adeline Nazarenko (Pr, Université de Paris 13, directrice), M. Emmanuel Morin (Pr, Université de Nantes, président), M. Patrice Bellot (MC, Université d'Avignon et des Pays de Vaucluse, rapporteur), M. Nabil Hathout (CR1, CLLE-CNRS, rapporteur), M. Olivier Gaunet (IR, Sinequa, examinateur), M. François Lévy (PR, Université de Paris 13, examinateur), M. Claude de Loupy (DR , Syllabs, examinateur).

***Résumé** : Les systèmes intégrant des traitements venant du traitement automatique des langues reposent souvent sur des lexiques et des grammaires, parfois indirectement sur des corpus. À cause de la quantité et de la complexité des informations qu'elles contiennent, ces ressources linguistiques deviennent facilement une source d'incohérence. Dans cette thèse, nous explorons les moyens d'améliorer la gestion des nombreuses ressources linguistiques d'un moteur de recherche industriel en dix-neuf langues qui fait appel à une analyse textuelle élaborée. Nous proposons une méthode pour formaliser l'architecture linguistique des traitements linguistiques et des ressources utilisées par ceux-ci. Cette formalisation explicite la façon dont les connaissances contenues dans les ressources sont exploitées. Grâce à elle, nous pouvons construire des outils de gestion qui respectent l'architecture du système. L'environnement ainsi mis en place se concentre sur la mise à jour et l'acquisition des ressources linguistiques, leur exploitation étant figée par des contraintes industrielles.*

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

Anne GARCIA-FERNANDEZ : (annegf@limsi.fr)

Titre : Génération de réponses en langue naturelle orales et écrites pour les systèmes de question-réponse en domaine ouvert

Mots-clés : question-réponse, génération de réponse en langue naturelle, acquisition de corpus.

***Title :** Generation of speech and written answers in natural language for question-answering systems in open domain*

***Keywords :** question-answering, natural language answer generation, corpus acquisition.*

Thèse de doctorat en Informatique, Université de Paris-Sud Orsay, LIMSI-CNRS, UFR d'Informatique, Orsaysous la codirection de Anne Vilnat (Pr, LIMSI-

CNRS & Université de Paris Sud) et Sophie Rosset (CR, LIMSI-CNRS). Thèse soutenue le 10/12/2010.

Jury : Mme Anne Vilnat (Pr, Université de Paris Sud, codirectrice), Mme Sophie Rosset (CR, LIMSI-CNRS, codirectrice), M. Daniel Luzzati (Pr, LIUM & Université du Maine, président), M. Frédéric Béchet (Pr, LIF & Université de la Méditerranée, rapporteur),M. Guy Lapalme (Pr, RALI-DIRO & Université de Montréal, rapporteur), M. Bernardo Magnini (Pr, FBK & Università di Trento, examinateur).

Résumé *: Les travaux présentés dans ce mémoire se situent dans le contexte de la réponse à une question. Contrairement à de nombreux travaux traitant de la recherche de l'information à fournir en réponse à une question, nous avons considéré la réponse à une question comme une interaction. Notre problématique principale a été de caractériser la forme, que peut prendre une réponse en interaction avec la question, qui puisse être produite par des systèmes de question-réponse. Nous exposons les enjeux de l'interaction en langue naturelle, puis, plus précisément, de l'interaction du type « réponse à une question » et ce, considérant deux modalités d'interaction : l'oral et l'écrit. Nous proposons une comparaison de réponses d'humains et de systèmes puis exposons des travaux montrant les caractéristiques d'une réponse prise comme appartenant à une interaction. Nous montrons que répondre n'est pas uniquement présenter une information mais fait partie d'une interaction entre deux locuteurs et exposons l'importance et l'intérêt pour les systèmes de question-réponse de produire une réponse en langue naturelle. Le terme réponse est précisé et le terme d'information-réponse est défini. Puis nous cherchons à spécifier ce que pourrait être une réponse pour les systèmes de question-réponse et constatons l'absence de corpus constitué de telles réponses. Afin de collecter un corpus de réponses à des questions, il est primordial d'étudier la forme des questions à poser.*

Nous présentons ainsi une étude de l'état de l'art sur les variations linguistiques des questions. Nous identifions les différents éléments qui peuvent varier dans une question intervenant dans une interaction avec une réponse. Ces éléments correspondent à des caractéristiques de la question elle-même (son type sémantique, sa syntaxe, l'interrogatif utilisé, le type de son verbe principal et l'objet sur lequel elle porte) mais aussi à des caractéristiques de la réponse attendue par la question (son type, sa nature unique ou multiple et sa valence).

Nous présentons ensuite la collecte du corpus de réponses. Elle a été organisée à l'écrit et à l'oral, auprès de plus de 150 locuteurs natifs du français. Les questions soumises aux participants ont été construites à partir des paramètres mis en valeur précédemment. Les objectifs fixés lors de la collecte du corpus sont explicités. Le protocole mis en œuvre, la construction du corpus de questions et le déroulement des passations sont décrits. Une évaluation du protocole est présentée et concerne notamment la spontanéité des réponses, la facilité de donner une réponse aux questions, l'obtention de réponses non succinctes et l'obtention de réponses sans information complémentaire.

Enfin, nous présentons différentes analyses faites sur le corpus cherchant à répondre à un ensemble de questions à se poser pour mettre en place un module de génération de réponses en langue naturelle dans un système de question-réponse. Ces questions sont : Est-il judicieux de réutiliser des éléments de la question ? Quelle information faut-il répondre ? Comment construire une réponse minimale ? La modalité (oral/écrit) influe-t-elle sur la formulation de réponse à générer ? Euh... et si on hésite ? Que répondre quand on n'a pas de réponse ? Comment répondre plusieurs informations-réponses ?
Les perspectives de ce travail sont multiples. Au-delà de poursuivre la collecte de corpus à partir d'un corpus plus large de questions, il serait intéressant de travailler sur des patrons de réponses moins restrictifs que ceux présentés dans ce document. De plus, en s'inspirant des résultats obtenus par l'analyse du corpus, nous souhaiterions implémenter un module de génération de réponses en interaction, au sein d'un système qui effectue une analyse linguistique de la question sur laquelle nous pourrions nous appuyer pour construire des réponses reprenant la question.
Reste bien évidemment la question de l'évaluation de la production de réponses en interaction. Nous savons combien l'évaluation d'un système est importante, notamment pour l'améliorer et le comparer à d'autres systèmes similaires, et une réflexion sur des méthodes et des mesures d'évaluation nous semble primordiale.

URL où la thèse pourra être téléchargée :
http://perso.limsi.fr/annegf/Recherche/pdf/Garcia-Fernandez_Anne_These2010.pdf

---

Olivier HAMON : (hamon@elda.org)

Titre : Vers une architecture générique et pérenne pour l'évaluation en traitement automatique des langues : spécifications, méthodologies et mesures

Mots-clés : évaluation, traitement automatique des langues, protocole d'évaluation, mesure d'évaluation, métrique d'évaluation, méta-évaluation, architecture d'évaluation, traduction automatique.

Title : *Towards a generic and sustainable architecture for evaluation in natural language processing : specifications, methodologies and measures*

Keywords : *evaluation, natural language processing, evaluation protocole, evaluation measure, evaluation metric, meta-evaluation, evaluation architecture, machine translation.*

Thèse de doctorat en Informatique, Université de Paris 13 – Paris Nord, Institut Galilée, LIPN – UMR CNRS 7030, Villetaneuse, sous la direction de Adeline Nazarenko (Pr, Université de Paris 13). Thèse soutenue le 06/12/2010.

Jury : Mme Adeline Nazarenko (Pr, Université de Paris 13, directrice), M. Joseph Mariani (DR, LIMSI-CNRS, président), M. Mohand Boughanem, (Pr, Université Paul-Sabatier, rapporteur), M. François Yvon (Pr, Université Paris 11, rapporteur), M. Khalid Choukri (PDG, ELDA, examinateur), M. Anthony Hartley (Pr, University of Leeds, examinateur), M. Daniel Kayser (Pr, Université Paris 13, examinateur).

Résumé *: Le développement de systèmes en traitement automatique des langues (TAL) nécessite de déterminer la qualité de ce qui est produit. Que ce soit pour comparer plusieurs systèmes entre eux ou identifier les points forts et faibles d'un système isolé, l'évaluation suppose de définir avec précision et pour chaque contexte particulier une méthodologie, un protocole, des ressources linguistiques (les données nécessaires à l'apprentissage et au test des systèmes) ou encore des mesures et métriques d'évaluation. C'est à cette condition que l'amélioration des systèmes est possible afin d'obtenir des résultats plus fiables et plus exploitables à l'usage. L'apport de l'évaluation en TAL est important avec la création de nouvelles ressources linguistiques, l'homogénéisation des formats des données utilisées ou la promotion d'une technologie ou d'un système.*

Toutefois, l'évaluation nécessite un important travail manuel, que ce soit pour l'expression des jugements humains ou pour la gestion du déroulement même de l'évaluation, ce qui compromet l'efficacité des évaluations, augmente leur coût et les rend difficilement reproductibles. Nous avons cherché à réduire et à encadrer ces interventions manuelles. Pour ce faire, nous appuyons nos travaux sur la conduite ou la participation à des campagnes d'évaluation comparant des systèmes entre eux, ou l'évaluation de systèmes isolés.

Nous avons formalisé la gestion du déroulement de l'évaluation et listé ses différentes phases pour définir un cadre d'évaluation commun, compréhensible par tous. Le point phare de ces phases d'évaluation concerne la mesure de la qualité via *l'utilisation de métriques. Cela a imposé trois études successives sur les mesures humaines, les mesures automatiques et les moyens d'automatiser le calcul de la qualité et enfin la méta-évaluation des mesures qui permet d'en évaluer la fiabilité. En parallèle, les mesures d'évaluation utilisent des ressources linguistiques dont les aspects pratiques et administratifs à travers les opérations de création, standardisation, validation, impact sur les résultats, coûts de production et d'utilisation, identification et négociation des droits doivent être pris en compte. Dans ce contexte, l'étude des similarités entre les technologies et entre leurs évaluations nous a permis d'observer les points communs et de les hiérarchiser. Nous avons montré qu'un petit ensemble de mesures permet de couvrir une large palette d'applications à des technologies distinctes.*

Notre objectif final était de définir une architecture d'évaluation générique, c'est-à-dire adaptable à tout type de technologie du TAL, et pérenne, c'est-à-dire permettant

la réutilisation de ressources linguistiques, mesures ou méthodes au cours du temps. Notre proposition se fait à partir des conclusions des étapes précédentes afin d'intégrer les phases d'évaluation à notre architecture et d'y incorporer les mesures d'évaluation, sans oublier la place relative à l'utilisation de ressources linguistiques. La définition de cette architecture s'est effectuée en vue d'automatiser entièrement la gestion des évaluations, que ce soit pour une campagne d'évaluation ou l'évaluation d'un système isolé. À partir de premières expérimentations, nous avons modélisé une architecture d'évaluation prenant en compte l'ensemble de ces contraintes et utilisant les services Web afin d'interconnecter les composants de l'architecture entre eux et d'y accéder via *le réseau Internet.*

URL où la thèse pourra être téléchargée :
Contacter l'auteur

Jean-François LUCAS : (jflucas@free.fr)

Titre : Modéliser le fonctionnement cognitif d'agents sociaux et reproduire des phénomènes linguistiques.

Mots-clés : agent intentionnel, agent rationnel, agent réactif, agent situé, agent social, automate fini, causalité, cognition, coopération, déterminisme, empirisme, enseignement, évolution, héritage, intentions, intelligence artificielle, jeu digital, langage objets, obstacles, phylogenèse, plate-forme, progression, séquenceur, simulation, système multi-agents, taxinomie, vie artificielle.

Title : *Modeling a cognitive functioning of social agents and reproducing some linguistic phenomena.*

Keywords : *artificial intelligence, artificial life, causality, cognition, cooperation, determinism, digital game, empiricism, evolution, finite automaton, intentional agent, inheriting, intentions, multi-agents system, object language, obstacles, phylogenesis, progression, rational agent, reactive agent, sequencer, simulation, situated agent, social agent, taxonomy, teaching, work-bench.*

Thèse de doctorat en Informatique, Université de Paris 8 – Saint-Denis, Institut Galilée, LIASD Laboratoire d'Informatique Avancée de Saint-Denis, UFR 6, sciences sociales, Saint-Denis, sous la direction de Gilles Bernard (Pr, Université de Paris 8). Thèse soutenue le 13/12/2010.

Jury : M. Gilles Bernard (Pr, Université de Paris 8, directeur), M. Patrick Greussay (Pr, Université de Paris 8, président), M. Herman Akdag (Pr, Université de Reims, rapporteur), M. Jacques Ferber, (Pr, Université Montpellier 2, rapporteur), M. Herbert Stoyan (Pr, Université d'Erlangen, rapporteur), M. Tristan Cazenave

(Pr, Université Paris-Dauphine, examinateur), M. Harald Wertz (Pr, Université Paris 8, examinateur).

Résumé *: Notre progression de simulations d'agents sociaux situés dans un monde digital modélise le fonctionnement cognitif d'agents coopératifs et reproduit des phénomènes linguistiques.*

Nous construisons pas à pas une progression d'agents situés en survie sur un support. Nous partons de la cellule végétative vue comme un séquenceur sur un chemin, puis enrichissons le support : nous plaçons des obstacles sur une grille. Pour ce faire, nous utilisons le modèle causal et déterministe de l'automate fini (AF), que nous simplifions en un agent réactif empirique. Ainsi, en formalisant la construction de l'agent situé (AS), nous décrivons un passage d'un domaine déterminisme à celui de l'empirisme. Puis nous introduisons la mémorisation, pour construire un agent intentionnel (AI) et continuons par quelques agents croyances désirs intentions (CDI). Finalement nous évoquons un peu les agents rationnels cognitifs (ARC).

Notre but : exhiber dans notre taxinomie des exemples de simulations significatives. Nous les construisons en guidant une émergence selon le trajet AF→AS→AI→Agent CDI→ARC. Dans cette progression, les agents évoluent en héritant de leurs parents. Ainsi ils se classent naturellement dans un arbre phylogénétique.

Cette plate-forme de recherche est aussi proposée pour l'enseignement. Le mécanisme interne des agents est écrit en C ; il est épuré, simple et accessible : il peut servir à enseigner l'informatique et l'IA au moyen d'exemples concrets. Notre outil s'inscrit dans le long terme : il est ouvert et peut être repris et prolongé par ceux qui voudront le faire évoluer.

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