TALN-RÉCITAL 2013

TALN : Traitement Automatique des Langues Naturelles

RÉCITAL : Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues

---

Actes de la conférence TALN-RÉCITAL 2013

Volume 3 : Ateliers

Éditeurs

Emmanuel Morin
Yannick Estève

17 au 21 juin 2013
Les Sables d'Olonne, France

Sous l'égide de l'ATALA (Association pour le Traitement Automatique des langues).

Avant-propos

Il est maintenant une tradition dans la communauté de l'ATALA de venir fouler tous les dix ans les côtes à l'ouest de la France. Ainsi après la Côte d'Amour en 2003, nous sommes heureux d'accueillir sur la Côte de Lumière la 20e conférence TALN et la 15e édition de RÉCITAL.

L'organisation de TALN et RÉCITAL 2013 a été assurée par les équipes TALN du LINA (Laboratoire d'Informatique de Nantes Atlantique) et LST du LIUM (Laboratoire d'Informatique de l'Université du Maine). Cette organisation conjointe est une bonne illustration de la synergie de ces deux équipes mais aussi de la dynamique du TALN dans la région des Pays de la Loire.

Cette année, avec 127 soumissions à TALN (dont 70 articles longs et 57 articles courts), la conférence a confirmé une fois encore son attractivité. Le processus d'évaluation, qui a demandé un travail important, a été réalisé consciencieusement pour arriver à une sélection de 36 articles longs et 35 articles courts. Nous remercions chaleureusement les membres des comités de lecture et de programme de TALN pour le travail réalisé. Outre ces communications, 13 démonstrations et 4 ateliers viennent accompagner la conférence. La conférence sera aussi ponctuée par l'intervention de deux conférenciers invités : Josiane Mothe et Alexander Fraser que nous tenons aussi à remercier de leur présence.

Poursuivant sur la démarche initiée lors de l'édition précédente, la conférence RÉCITAL a ciblé un large spectre de publications (états de l'art, travaux préliminaires, etc.). L'accent a une nouvelle fois été mis sur la pédagogie et l'échange direct en fournissant des relectures explicatives et non anonymes aux auteurs. Cette formule fonctionne bien puisque l'on recense un total de 25 soumissions parmi lesquelles 18 ont été sélectionnées (6 présentations orales et 12 posters). Nous remercions les membres du comité de programme de RÉCITAL pour les précieux retours qui, nous en sommes convaincus, sont très appréciés des jeunes chercheurs en TALN.

Cette conférence sur un site distant, mais presque à mi-chemin entre Le Mans et Nantes, a nécessité un travail d'organisation important. Que les membres du comité d'organisation trouvent ici la reconnaissance du travail réalisé.

Nous n'oublions pas non plus les partenaires institutionnels et privés qui se sont joints à nous pour faire de cette 20e conférence TALN et de cette 15e édition de RÉCITAL une véritable réussite.

Comme il est de tradition aux Sables d'Olonne, nous souhaitons à l'ensemble des conférenciers « bon vent ».

Emmanuel Morin
Yannick Estève
Organisateurs de TALN 2013

Florian Boudin
Loïc Barrault
Organisateurs de RÉCITAL 2013

Comité d'organisation de TALN-RÉCITAL

Président de TALN

Emmanuel Morin	LINA	Université de Nantes, France

Vice-Président de TALN

Yannick Estève	LIUM	Université du Maine, France

Présidents de RÉCITAL

Florian Boudin	LINA	Université de Nantes
Loïc Barrault	LIUM	Université du Maine

Membres

Denis Béchet	LINA	Université de Nantes
Fethi Bougares	LIUM	Université du Maine
Adrien Bougouin	LINA	Université de Nantes
Nathalie Camelin	LIUM	Université du Maine
Béatrice Daille	LINA	Université de Nantes
Colin De La Higuera	LINA	Université de Nantes
Paul Deléglise	LIUM	Université du Maine
Estelle Delpech	LINA	Université de Nantes
Alexandre Dikovsky	LINA	Université de Nantes
Chantal Enguehard	LINA	Université de Nantes
Rima Harastani	LINA	Université de Nantes
Mohamed Hatmi	LINA	Université de Nantes
Amir Hazem	LINA	Université de Nantes
Nicolas Hernandez	LINA	Université de Nantes
Firas Hmida	LINA	Université de Nantes
Christine Jacquin	LINA	Université de Nantes
Ophelie Lacroix	LINA	Université de Nantes
Antoine Laurent	LIUM	Université du Maine
Elizaveta Loginova	LINA	Université de Nantes
Daniel Luzzati	LIUM	Université du Maine
Sylvain Meignier	LIUM	Université du Maine
Laura Monceaux	LINA	Université de Nantes
Simon Petitrenaud	LIUM	Université du Maine
Emmanuel Planas	LINA	Université Catholique de l'Ouest
Solen Quiniou	LINA	Université de Nantes
Anne-Françoise Quin	LINA	Université de Nantes
Holger Schwenk	LIUM	Université du Maine
James Scicluna	LINA	Université de Nantes
Christophe Servan	LIUM	Université du Maine
Prajol Shrestha	LINA	Université de Nantes
Déborah Sourdillat	LINA	Université de Nantes
Annie Tartier	LINA	Université de Nantes

Atelier DEFT 2013 : neuvième Défi Fouille de Texte

Comité d'organisation

Flora Badin	INIST, Nancy
Dominic Forest	EBSI, U. Montréal
Cyril Grouin	LIMSI-CNRS, Orsay
Patrick Paroubek	LIMSI-CNRS, Orsay
Pierre Zweigenbaum	LIMSI-CNRS, Orsay

Comité de programme

Flora Badin	INIST, Nancy
Nathalie Camelin	LIUM, Le Mans
Béatrice Daille	LINA, Nantes
Dominic Forest	EBSI, U. Montréal
Cyril Grouin	LIMSI-CNRS, Orsay
Patrick Paroubek	LIMSI-CNRS, Orsay
Pierre Zweigenbaum	LIMSI-CNRS, Orsay

Atelier MIXEUR 2013 : méthodes mixtes pour l'analyse syntaxique et sémantique du français

Responsable

Christian Retoré	Université Bordeaux, IRIT et LaBRI

Comité de programme

Laurence Danlos	Université Paris Diderot et ALPAGE
Richard Moot	CNRS LaBRI
Jean-Philippe Prost	Université Montpellier II et LIRMM
Tim Van de Cruys	CNRS IRIT

Atelier SemDis 2013 : enjeux actuels de la sémantique distributionnelle

Comité d'organisation

Cécile Fabre	CLLE, Toulouse
Nabil Hathout	CLLE, Toulouse
Philippe Muller	IRIT, Toulouse
Tim Van de Cruys	IRIT, Toulouse

Comité de programme

Stergos Afantenos	IRIT, Toulouse
Yves Bestgen	UCL/CECL, Louvain-La-Neuve
Marie Candito	ALPAGE, Paris
Eric de la Clergerie	ALPAGE, Paris
Cécile Fabre	CLLE, Toulouse
Olivier Ferret	CEA-LIST, Fontenay-aux-Roses
Nabil Hathout	CLLE, Toulouse
Philippe Muller	IRIT, Toulouse
Adeline Nazarenko	LIPN, Paris
Pascale Sébillot	IRISA, Rennes
Ludovic Tanguy	CLLE, Toulouse
Agnès Tutin	LIDILEM, Grenoble
Tim Van de Cruys	IRIT, Toulouse
Virginie Zampa	LIDILEM, Grenoble

Atelier TALaRE 2013 : Traitement Automatique des Langues Régionales de France et d'Europe

Comité d'organisation

Marianne Vergez-Couret	CLLE-ERSS, Université de Toulouse 2
Delphine Bernhard	LILPA, Université de Strasbourg
Jean-Michel Eloy	LESCLAP, Université de Picardie
Christophe Rey	LESCLAP, Université de Picardie

Comité de programme

Martine Adda-Decker	LPP, Université Paris 3
Vincent Berment	CLIPS-GÉTA, Université Joseph Fourier, Grenoble
Myriam Bras	CLLE-ERSS, Université de Toulouse 2
Alain Dawson	LESCLAP, Université de Picardie
Nuria Gala	LIF, Aix-Marseille Université
Nabil Hathout	CLLE-ERSS, Université de Toulouse 2
Anne-Laure Ligozat	LIMSI, CNRS
Jean-Marie Pierrel	ATILF, Université de Lorraine
Yves Scherrer	ALPAGE, Université Paris 7
Pascal Vaillant	Université Paris 13

Partenaires

Or

Argent

Bronze

Institutionnels

Table des matières

Actes de DEFT 2013 : 9e DÉfi Fouille de Textes 1

DEFT2013 se met à table : présentation du défi et résultats
Cyril Grouin, Pierre Zweigenbaum et Patrick Paroubek 1

Efficacité combinée du flou et de l'exact des recettes de cuisine
Thierry Hamon, Amandine Périnet et Natalia Grabar 15

DEFT2013, une cuisine de caractères
Gaël Lejeune, Charlotte Lecluze et Romain Brixtel 29

Systèmes du LIA à DEFT'13
Xavier Bost, Ilaria Brunetti, Luis Adrián Cabrera-Diego, Jean-Valère Cossu, Andréa Linhares, Mohamed Morchid, Juan-Manuel Torres-Moreno, Marc El-Bèze et Richard Dufour 37

Approches hybrides pour l'analyse de recettes cuisine DEFT, TALN-RECITAL 2013
Luca Dini, André Bittar et Mathieu Ruhlmann 48

Participation d'Orange Labs à DEFT2013
Olivier Collin, Aleksandra Guerraz, Yannick Hiou et Nicolas Voisine 61

Trois recettes d'apprentissage automatique pour un système d'extraction d'information et de classification de recettes de cuisine
Eric Charton, Ludovic Jean-Louis, Marie-Jean Meurs et Michel Gagnon 75

Actes de MIXEUR 2003 : Méthodes mixtes pour l'analyse syntaxique et sémantique du français 88

Clustering pour grammaires catégorielles du second ordre
Noémie-Fleur Sandillon-Rezer 88

Des bigrammes au secours de la grammaire
Eric Wehrli, Luka Nerima et Meghdad Farahmand 92

Intégrer des ressources lexicales et grammaticales externes dans des analyseurs partiels probabilistes
Matthieu Constant et Isabelle Tellier 95

Compositionnalité en sémantique distributionnelle et ordonnancement des adjectifs
Eva Maria Vecchi, Roberto Zamparelli et Marco Baroni 99

Inférence de règles déductives par abduction
Mathieu Lafourcade, Manel Zarrouk et Alain Joubert 103

Probabiliser les Grammaires de Propriétés
Philippe Blache et Stéphane Rauzy 108

Actes de SemDis 2013 : Enjeux actuels de la sémantique distributionnelle 112

Utiliser une base distributionnelle pour filtrer un dictionnaire de synonymes
François Morlane-Hondère 112

Exploration sémantique visuelle à partir des cooccurrences de deuxième et troisième ordre
Ann Bertels et Dirk Speelman 126

Étude des relations sémantiques dans les reformulations de requêtes sous la loupe de l'analyse distributionnelle
Clémentine Adam, Cécile Fabre et Ludovic Tanguy 140

Visualisations interactives des espaces vectoriels sémantiques pour l'analyse lexicologique
Thomas Wielfaert, Kris Heylen et Dirk Speelman 154

Vers un environnement de production et de validation de ressources lexicales sémantiques
Mikaël Morardo et Éric Villemonte de La Clergerie 167

Actes de TALARE 2013 : Traitement automatique des langues régionales de France et d'Europe 181

Typologie et ressources pour le TALN des langues caucasiennes : le cas du tchétchène
Johannes Heinecke 181

Étiquetage morphosyntaxique de langues non dotées à partir de ressources pour une langue étymologiquement proche
Yves Scherrer et Benoît Sagot 195

Es esch fàscht wie Ditsch, oder net ? Étiquetage morphosyntaxique de l'alsacien en passant par l'allemand
Delphine Bernhard et Anne-Laure Ligozat 209

Jochre, océrisation par apprentissage automatique : étude comparée sur le yiddish et l'occitan
Assaf Urieli et Marianne Vergez-Couret 221

A la recherche de Technologies de la Langue pour les langues minoritaires
Claudia Soria et Joseph Mariani 235

Liste des auteurs 248

Liste des mots clés 250

DEFT2013 se met à table : présentation du défi et résultats

Cyril Grouin[1,2] Pierre Zweigenbaum[1] Patrick Paroubek[1]
(1) LIMSI–CNRS, Orsay (2) INSERM U872 Eq 20 & UPMC, Paris
prenom.nom@limsi.fr

RÉSUMÉ
L'édition 2013 du défi fouille de texte (DEFT) a porté sur l'analyse des recettes de cuisine sous quatre angles différents : le niveau de difficulté d'une recette, le type de plat préparé, le titre, et les ingrédients nécessaires à la réalisation de la recette. Dans cet article, nous présentons le déroulement du défi au travers des corpus produits et des tests humains réalisés, ainsi que les résultats obtenus par les équipes ayant participé. Les mesures d'évaluation utilisées sont la distance relative moyenne à la solution pour la difficulté du plat et le couple précision/rappel pour le type de plat, la MRR (moyenne de l'inverse du rang) pour l'association titre/recette et la MAP (moyenne des précisions moyennes) pour la liste des ingrédients d'une recette.

ABSTRACT
DEFT2013 spills the beans: challenge presentation and results

The DEFT 2013 challenge focused on the analysis of cooking recipes through four points of view: the level of difficulty of the recipe, the type of dish to be prepared, the title, and the ingredients needed to prepare the recipe. In this paper, we present the challenge: the corpora we produced, the human tests, as well as the results each participant achieved. The evaluation measures we used are the mean distance to the solution for the difficulty task, precision and recall for the type of dish, mean reciprocal rank (MRR) for the title/recipe association and mean average precision (MAP) for the list of ingredients.

MOTS-CLÉS : campagne d'évaluation, classification, mesures, recettes.

KEYWORDS: evaluation campaign, classification, metrics, recipes.

1 Présentation

1.1 Introduction

Le défi DEFT est un atelier annuel d'évaluation francophone en fouille de textes. La neuvième édition de ce défi a porté sur l'analyse des recettes de cuisine rédigées en langue française. Dans le cadre de cette nouvelle édition, nous avons orienté le défi vers un nouveau domaine d'application. La thématique retenue, les recettes de cuisine, a déjà fait par le passé l'objet d'une campagne d'évaluation *(Computer Cooking Contest*[1]*)*. Nous nous intéressons dans DEFT2013 à deux types de fonction d'analyse du langage, la classification de documents *(tâches 1 à 3)* et l'extraction d'information *(tâche 4)*, ceci dans un domaine de spécialité.

1. http://computercookingcontest.net

 © ATALA

Pour cette nouvelle édition du défi, nous proposons quatre tâches d'analyse concernant les recettes de cuisine : identifier le niveau de difficulté de réalisation d'une recette, identifier le type de plat préparé, apparier une recette à son titre, identifier les ingrédients d'une recette.

1.2 Organisation

Une étape de réflexion relative au choix des thématiques et des tâches à proposer a été engagée en fin d'année 2012. Durant cette période, des tests de tâches ont été effectués auprès d'étudiants. Ces tests nous ont permis de confirmer les choix effectués, ou dans certains cas, de réorienter certaines tâches envisagées. Nous avons orienté cette nouvelle édition vers les recettes de cuisine du site Marmiton, [2] dont les catégories d'informations proposées correspondaient aux thématiques que nous souhaitions traiter dans ce défi et parce que les recettes sont librement accessibles.

Deux appels à participation ont été lancés sur la liste de diffusion de la communauté du traitement automatique des langues LN les 16 février et 5 mars. Les données d'apprentissage ont été mises à la disposition des participants à partir du 28 février, les données de test l'ont été entre le 25 avril et le 5 mai, dans une fenêtre de trois jours au libre choix de chaque participant. La référence et les résultats individuels *(avec classement, moyenne et médiane)* ont été communiqués aux participants le 6 mai.

Dix équipes se sont inscrites à cette édition du défi, parmi lesquelles deux sont issues du monde industriel. Les six équipes qui ont participé aux tests sont les suivantes :
- **Celi France :** Luca Dini, André Bittar, Mathieu Ruhlmann ;
- **GREYC :** Gaël Lejeune, Charlotte Lecluze, Romain Brixtel ;
- **LIA :** Xavier Bost, Ilaria Brunetti, Luis Adrián Cabrera-Diego, Jean-Valère Cossu, Andréa Linhares, Mohamed Morchid, Juan-Manuel Torres-Moreno, Marc El Bèze, Richard Dufour ;
- **LIM&Bio :** Thierry Hamon, Natalia Grabar, Amandine Périnet ;
- **Orange Labs :** Olivier Collin, Aleksandra Guerraz, Yannick Hiou, Nicolas Voisine ;
- **Wikimeta Lab :** Eric Charton, Ludovic Jean-Louis, Marie-Jean Meurs, Michel Gagnon.

2 Présentation

2.1 Tâches proposées

Nous avons proposé quatre tâches autour de la thématique des recettes de cuisine. La motivation des deux premières tâches concerne le jugement appliqué à une recette. Dans de nombreux cas, le type de plat, et plus encore, le niveau de difficulté de réalisation d'une recette peut se révéler variable d'un individu à un autre, avec pour conséquence une indexation biaisée des recettes présentes sur le site. L'indexation des recettes selon les ingrédients qu'elle contient est également une tâche complexe, qui fait appel à la distinction entre ingrédients obligatoires et ingrédients facultatifs. Enfin, l'appariement d'un titre avec la recette qui lui correspond permet de mettre en évidence les désaccords qui peuvent parfois exister dans ces associations. L'objectif global visé par les quatre tâches du défi vise ainsi à fournir des méthodes adaptées à l'indexation de sites web participatifs, dont le contenu, déposé par chaque utilisateur, est généralement classé et

2. http://www.marmiton.org/

© ATALA

évalué par l'utilisateur qui aura déposé le contenu en ligne, avec des écarts d'appréciation de classement assez importants.

Tâche 1 — Niveau de difficulté de réalisation d'une recette. La première tâche a pour but d'évaluer la capacité d'un algorithme à inférer la difficulté d'une recette de cuisine en se basant sur toutes les informations qu'il est possible d'extraire à partir du texte de la recette et de son titre. Le niveau de difficulté de réalisation d'une recette est évalué sur une échelle à quatre valeurs : *très facile, facile, moyennement difficile, difficile*. La mesure d'évaluation principale retenue pour cette tâche est la distance moyenne à la solution, calculée en micro-moyenne sur une échelle à quatre valeurs. Nous avons aussi vérifié la corrélation des résultats ainsi obtenus avec ceux provenant d'une mesure de précision/rappel en micro-mesure.

Tâche 2 — Type de plat. La deuxième tâche propose de classer les recettes en fonction du type de plat préparé, selon une partition en trois classes : *entrée, plat principal, dessert*. Les mesures d'évaluation retenues sont la précision et le rappel calculées en micro-moyennes.

Tâche 3 — Appariement titre/recette. La troisième tâche demande au système de retrouver pour chaque texte de recette à traiter, son titre original dans une liste de titres de recettes.

Tâche 4 — Ingrédients d'une recette. La dernière tâche se démarque des précédentes car elle ne concerne pas la classification des recettes, mais l'extraction d'information. Il s'agit en effet dans cette tâche d'identifier la liste des ingrédients de la recette. Les ingrédients identifiés doivent être exprimés selon des libellés normalisés présents dans une liste globale fournie aux participants. Cette liste contient au moins tous les ingrédients à trouver dans la base des textes de recettes, mais peut aussi contenir des ingrédients qui ne sont présents dans aucune des recettes des corpus d'entraînement et de test.

2.2 Corpus

2.2.1 Présentation

Le corpus de recettes utilisées pour le défi provient du site de recettes participatif Marmiton. Sur ce site, chaque utilisateur dépose ses propres recettes. Ce type de dépôt implique que l'utilisateur renseigne lui-même certaines des caractéristiques liées à la recette, en particulier les informations de coût et de niveau de difficulté. Les participants ont été autorisés à utiliser des corpus de recettes complémentaire, à l'exclusion des recettes provenant de Marmiton.

2.2.2 Modalités de constitution

Collecte des données. Nous avons téléchargé les recettes du site pendant les deux premières semaines de février 2013, au moyen d'une chaîne effectuant les étapes suivantes :
- Collecte des 26 pages d'index des ingrédients utilisés dans les recettes du site ;
- Pour chaque ingrédient indexé dans les pages d'index précédentes, collecte des pages listant les recettes utilisant cet ingrédient ;
- Conversion des pages web téléchargées au format XML.

© ATALA

Au terme de cette phase de récupération, nous avons collecté 46 176 recettes que nous avons transformées en XML. La moitié de ces recettes a été aléatoirement conservée pour l'édition 2013 du défi, l'autre moitié pouvant servir de corpus pour l'année suivante.

Constitution de la référence. La référence des différentes tâches a été réalisée de manière automatique, sur la base des informations présentes dans les recettes.

– la référence des tâches 1 à 3 est directement inférée des informations présentes dans les recettes : *niveau de difficulté, type de plat, titre de la recette* ;
– la référence de la tâche 4 a été constituée lors de la collecte des données par la correspondance entre les pages d'index des ingrédients et les différentes recettes.

Concernant les trois premières tâches, la référence est issue des informations renseignées par l'utilisateur qui a posté la recette sur Marmiton. Ainsi, le niveau de difficulté de réalisation d'une recette est largement subjectif et dépend également du niveau d'expertise de chacun en cuisine ; d'autre part, l'auteur d'une recette aura tendance à minorer le niveau de difficulté de sa recette, de manière à inciter le plus grand nombre de visiteurs à tester la recette avec l'espoir d'obtenir des commentaires en retour. On observe ce comportement dans la répartition des recettes selon les quatre niveaux de difficultés (tableau 1 gauche). À l'inverse, la répartition en types de plat semble plus équilibrée et moins sujette à la subjectivité (tableau 1 droite). Nous sommes conscients de la présence d'erreurs dans les références mais nous avons considéré que ces erreurs étaient marginales et n'auraient pas un impact trop fort sur l'évaluation finale. En conséquence, aucune phase de nettoyage humain *(tâche forcément coûteuse)* n'a été réalisée sur les références.

Niveau	Nombre	Pourcentage	Type	Nombre	Pourcentage
Très facile	11 602	50,2 %	Entrée	5 407	23,4 %
Facile	9 584	41,5 %	Plat principal	10 742	46,5 %
Moyennement difficile	1 776	7,7 %	Dessert	6 945	30,1 %
Difficile	132	0,6 %			

TABLE 1 – Répartition des recettes par niveau de difficulté (gauche) et type de plat (droite)

En ce qui concerne la dernière tâche, chaque recette contient la liste des ingrédients fournie par l'utilisateur. Cependant, les ingrédients de cette liste ont une forme qui varie avec les recettes, et il serait très difficile d'arriver à prédire exactement le libellé utilisé par l'auteur. C'est pourquoi la référence a été constituée à partir de l'index des ingrédients des recettes, index qui constitue une liste de libellés normalisés d'ingrédients et qui associe à chacun l'ensemble des recettes qui l'utilisent. Si cette méthode a l'avantage de fournir une liste normalisée d'ingrédients pour chaque recette, elle présente néanmoins plusieurs défauts. D'une part, certains ingrédients d'une recette ont pu être oubliés lors de l'indexation, et seront considérés comme faux positifs s'ils sont (correctement) trouvés par un système. D'autre part, certains index peuvent être associés par erreur à une recette (c'est le cas du « maïs » mis comme index d'une recette dont un ingrédient contenait l'expression « mais compter un peu plus long pour la cuisson »), et ils compteront comme faux négatifs si un système ne les produit pas. À noter, la forme normalisée des ingrédients ne contient pas de caractères accentués (les diacritiques sont supprimés) ni d'espaces (ils sont remplacés par des tirets).

 © ATALA

Répartition des données. Le corpus de l'édition 2013 totalise 23 094 recettes que nous avons ensuite réparties entre corpus d'apprentissage *(13 864 recettes)* et corpus de test *(9 230 recettes)* selon une répartition respectivement fixée à 60 %/40 % comme utilisée dans les précédentes campagnes de DEFT. Cette répartition a été effectuée de telle sorte que la proportion de recettes en termes de niveau de difficulté et de type de plat soit équivalente dans les deux corpus.

Le corpus d'apprentissage est commun à l'ensemble des quatre tâches proposées cette année, dans la mesure où les données de référence sont incluses dans les méta-informations. En ce qui concerne le corpus de test, nous l'avons segmenté en quatre parties égales *(soit environ 2 300 documents par partie)*, chaque partie étant réservée pour une tâche, en garantissant également que chaque partie conserve la même proportion de recettes en niveaux de difficulté et type de plat *(même si cet impératif ne se révèle nécessaire que pour les tâches 1 et 2)*. Il en ressort que les documents des corpus de test de chacune des quatre tâches sont distincts pour chaque tâche.

Formatage des données. Nous donnons en figure 1 un exemple de recette issue du corpus d'apprentissage. Chaque document se compose de deux parties principales :
- des méta-données contenant : le titre *(référence de la tâche 3)*, le type de plat *(référence de la tâche 2)*, le niveau de difficulté *(référence de la tâche 1)*, le coût *(information complémentaire fournie dans l'apprentissage et le test)*, et la liste des ingrédients renseignés par l'utilisateur ;
- la préparation de la recette contenant les différentes étapes.

En dehors du document, la liste des ingrédients normalisés *(référence de la tâche 4)* a été fournie. La liste des ingrédients normalisés qui indexent cette recette sont les suivants : *ail, basilic, bouillon, citron, echalote, huile-d-olive, parmesan, pignon, poivre, tomate, truite*. Notons que les ingrédients normalisés, parce qu'ils servent à indexer le contenu du site Marmiton, étaient déjà tous désaccentués, avec espaces et apostrophes remplacés par des tirets.

2.3 Tests humains

Nous avons mis à contribution les étudiants de l'EBSI [3] et du M2 Pro d'Ingénierie Linguistique de l'INaLCO [4] en novembre 2012. Les étudiants ont travaillé sur les quatre tâches que nous envisagions à l'époque : (i) évaluer le niveau de difficulté *(très facile, facile, moyennement difficile, difficile)* d'une recette ; (ii) prédire le type de plat *(entrée, plat, dessert)* ; (iii) relier un titre à la recette correspondante ; et (iv) identifier l'ingrédient ajouté à la recette d'origine.

Niveau de difficulté. Les étudiants de l'INaLCO ont pris une demi-heure pour étudier le corpus et prédire le niveau de difficulté. Les résultats de ces tests sont rassemblés dans le tableau 2. La bonne réponse *(niveau de difficulté renseigné dans Marmiton)* est identifiée, au minimum par aucun annotateur et au maximum par 7 annotateurs, avec un taux moyen de réussite sur l'ensemble du corpus de 37,0 %. Aucune recette n'est donc évaluée de la même manière par l'ensemble des annotateurs. La majorité des erreurs effectuées (57,1 %) ne correspond toutefois

3. Ecole de Bibliothéconomie et des Sciences de l'Information, Université de Montréal. Cours assuré par Dominic Forest auprès d'un groupe de douze étudiants.

4. Institut National des Langues et Civilisations Orientales, Paris. Cours assuré par Cyril Grouin pour les parcours « *Ingénierie Multilingue* » et « *Traductiques et gestion de l'information* » du M2 Pro d'Ingénierie Linguistique auprès d'un groupe de dix étudiants : Abdennour Goumiri, Ardas Khalsa, Arthur Boyer, Asma Benahmed, Hamza Affane, Julie Arnal, Laure Chancerelle, Maria Goryainova, Selen Şahin et Thomas Moraine. Tous ces étudiants se sont acquittés de cette tâche d'évaluation humaine avec abnégation, à quelques heures seulement du dîner... Nous les en remercions vivement.

© ATALA

```
<?xml version="1.0" encoding="utf-8"?>
<recette id="54562">
  <titre>Cuisses de poulet au miel cuites au four</titre>
  <type>Plat principal</type>
  <niveau>Très facile</niveau>
  <cout>Bon marché</cout>
  <ingredients>
    <p>2 cuisses de poulet</p>
    <p>1 oignon</p>
    <p>10 cl d'huile d'olive (environ 1/2 verre à eau)</p>
    <p>3 cuillères à soupe de miel</p>
    <p>sel, poivre</p>
  </ingredients>
  <preparation>
  <![CDATA[
    Préchauffez le four à 200'C (thermostat 6-7).
    Peler les oignons et les émincer.
    Les disposer dans un plat à gratin (environ 40 x 20 cm).
    Disposer les cuisses de poulet sur les oignons.
    Dans un bol, mélangez l'huile et le miel.
    Badigeonner le poulet de ce mélange.
    Saler et poiver les cuisses.
    Mettre le plat dans le four pendant environ 50 minutes.
    Vos oignons sont caramélisés et de même pour vos cuisses de poulet.
    Bonne dégustation.
    Cette recette accompagnée par une bonne purée, c'est simple et c'est un vrai
        régal!Vous pouvez ajouter un deuxième oignon selon vos envies.
  ]]>
  </preparation>
</recette>
```

FIGURE 1 – Recette issue du corpus d'apprentissage avec la référence des différentes pistes en méta-données (sauf pour la tâche 4, dont les ingrédients normalisés pour cette recette étaient *cuisse-de-poulet, huile-d-olive, miel, oignon, poivre, sel*)

qu'à des écarts d'un seul niveau. Enfin, les annotateurs humains parviennent à identifier si la recette est plutôt d'un niveau facile ou d'un niveau difficile. La même expérience réalisée sur un second corpus de dix recettes auprès d'un groupe d'étudiants de l'EBSI permet l'obtention de résultats similaires avec un taux de réussite moyen de 34,0 %.

D'autre part, les annotateurs sont majoritairement d'accord entre eux. Si l'évaluation est réalisée non plus d'après le niveau de référence mais selon le niveau majoritairement attribué par les annotateurs, le taux moyen de réussite sur l'ensemble du corpus monte à 52,0 %. Cette observation témoigne du fait que le niveau de difficulté d'origine est mal choisi pour un grand nombre de recettes. Les concepteurs de sites participatifs qui évaluent la qualité d'un produit *(recettes, livres, films, etc.)* pourraient donc être intéressés par des outils d'évaluation automatique.

Type de plat. La tâche de prédiction du type de plat *(entrée, plat, dessert)* a été réalisée par les douze étudiants de l'EBSI. Sur cette tâche, ils ont fourni entre 6 et 9 bonnes réponses sur le corpus de dix recettes, avec un taux moyen de réussite de 77,5 %. À l'évidence, la tâche semble moins simple qu'il n'y paraît, a fortiori sur la distinction entre une entrée et un plat principal.

 © ATALA

Recette	Référence	Annotateur									
		01	02	03	04	05	06	07	08	09	10
Couscous végétarien	1	4	3	**1**	3	3	3	2	3	4	3
Pain d'épice	1	2	2	**1**	**1**	**1**	**1**	2	3	2	**1**
Tartiflette courgettes	1	**1**	**1**	**1**	2	**1**	**1**	2	2	**1**	**1**
Garbure landaise	2	3	4	1	4	4	**2**	**2**	3	4	3
Rougets farcis	2	3	3	4	**2**	**2**	4	**2**	3	4	4
Lasagnes bolognaise	3	2	**3**	2	2	2	**3**	1	**3**	2	**3**
Sushi californien	3	**3**	4	**3**	**3**	**3**	4	**3**	2	4	**3**
Boulettes orientales	4	1	2	2	3	1	2	3	3	3	1
Lasagnes de légumes	4	3	2	**4**	3	3	**4**	2	**4**	3	3
Ravioles tartares	4	**4**	**4**	**4**	0	3	3	3	**4**	**4**	**4**

TABLE 2 – Prédictions des annotateurs humains en matière de niveau de difficulté d'une recette (0 : absence de réponse, 1 : très facile, 2 : facile, 3 : moyennement difficile, 4 : difficile)

Appariement titre/recette. La tâche consistant à relier chaque recette à son titre *(un corpus de dix recettes et de dix titres séparés)* a été réalisée par les étudiants de l'EBSI qui ont pratiquement tous réussi à effectuer les appariements corrects ; seul un(e) étudiant(e) a interverti deux titres. Le taux moyen de réussite est donc élevé sur cette tâche et se monte à 98,3 %. Le passage à l'échelle avec plusieurs milliers de titres et de recettes à apparier devrait se révéler plus complexe.

Ingrédient étranger. Sur la tâche d'identification de l'ingrédient étranger à une recette, [5] tous les étudiants de l'INaLCO ont correctement identifié l'ingrédient ajouté. Pour le défi, nous avons remplacé cette tâche par l'identification des ingrédients normalisés qui constituent la recette.

3 Évaluation et discussion

3.1 Modalités d'évaluation

Tâches 1 & 2 Les étiquettes de classe de la tâche 1 correspondent à des graduations sur une échelle de valeurs. Nous avons donc retenu comme mesure principale la distance relative moyenne à la solution, calculée en micro-moyenne comme mesure principale et en macro-moyenne comme mesure secondaire. L'échelle de valeurs comprend 4 valeurs réparties de manière symétrique sur une droite euclidienne (un espace à une dimension), de part et d'autre de l'origine. La distance entre les deux premières valeurs de classe de difficultés opposées est choisie comme double de celle séparant deux valeurs de difficulté de même signe (par ex. d(*très facile*, *facile*)= 1 mais d(*facile*, *moyennement difficile*)= 2), ceci afin de pénaliser plus un système qui ferait une erreur de genre de difficulté (facile/moyennement difficile) qu'un système faisant une erreur de qualification de difficulté (*très facile*/*facile*). Comme mesure secondaire, nous considérons le même calcul en macro-moyenne, afin de séparer des systèmes potentiellement ex-aequo en observant de manière plus fine leur performance sur les classes de recettes ayant des effectifs de

5. Pour chacune des cinq recettes constituant le corpus, nous avons ajouté un ingrédient provenant d'une recette du même type de plat. Sur la recette *« Crevettes au riz et à l'avocat en coque de tomate »* (entrée), nous avons ajouté des tomates cerises, tandis que sur la recette *« Cheesecake à la carotte »* (dessert) nous avons ajouté du sirop d'érable.

© ATALA

petite taille. À chacune des 4 valeurs possibles pour la donnée de référence r_i, correspond une valeur de distance maximale possible entre la réponse du système et cette donnée $dmax(h_i, r_i)$. Par exemple si la bonne réponse est *très facile*, la distance maximale possible est 4 si la réponse du système est *difficile*, mais si la bonne réponse est *facile*, la distance maximale possible sera 3, pour une réponse du système *difficile*. L'exactitude en distance relative à la solution moyenne (EDRM) se calcule en micro-moyenne comme indiqué dans l'équation 1.

$$\text{EDRM} = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{d(h_i, r_i)}{dmax(h_i, r_i)}\right) \tag{1}$$

Cependant, la première tâche du défi peut aussi être considérée comme une tâche de classification automatique, comme l'est la tâche 2. À ce titre, les mesures habituelles de rappel, précision et F-mesure sont donc adaptées. Nous avons donc vérifié la corrélation des résultats obtenus avec l'EDRM sur la tâche 1 avec les mesures de précision et rappel, qui sont les mesures principales de la tâche 2. Ces mesures sont calculées à la fois en termes de micro-mesure (formules 2 et 3) et de macro-mesures (formules 4 et 5), le classement officiel étant établi sur la base des micro-mesures de manière à accorder un poids équivalent à chaque élément mesuré, même si cette mesure revient à privilégier les classes composées d'un grand nombre d'individus *(très facile, facile)* au détriment des classes faiblement représentées *(moyennement difficile, difficile)*.

$$\text{Micro-rappel} = \frac{\sum_{i=1}^{n} \text{vrais positifs}(i)}{\sum_{i=1}^{n} \text{vrais positifs}(i) + \sum_{i=1}^{n} \text{faux négatifs}(i)} \tag{2}$$

$$\text{Micro-précision} = \frac{\sum_{i=1}^{n} \text{vrais positifs}(i)}{\sum_{i=1}^{n} \text{vrais positifs}(i) + \sum_{i=1}^{n} \text{faux positifs}(i)} \tag{3}$$

$$\text{Macro-rappel} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\text{vrais positifs}(i)}{\text{vrais positifs}(i) + \text{faux négatifs}(i)}\right) \tag{4}$$

$$\text{Macro-précision} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\text{vrais positifs}(i)}{\text{vrais positifs}(i) + \text{faux positifs}(i)}\right) \tag{5}$$

Tâches 3 et 4. Les tâches 3 et 4 demandaient de renvoyer une liste ordonnée d'hypothèses, l'objectif étant de placer le plus haut possible dans cette liste la ou les bonnes hypothèses, et peuvent de ce fait être mises en parallèle avec le format et le mode d'évaluation des tâches de recherche d'information. Dans la tâche 3, une seule bonne réponse (le titre original de la recette) était possible pour chaque recette R_i. Une mesure basée sur le rang r_i de cette bonne réponse a donc été employée : l'inverse de ce rang, et donc pour l'ensemble des N recettes la moyenne de l'inverse du rang (MRR, voir formule 6). Dans la tâche 4, plusieurs bonnes réponses (les n_i ingrédients $\{I_i^1 \ldots I_i^j \ldots I_i^{n_i}\}$ de la recette R_i) étaient attendues pour chaque recette. La moyenne des précisions non interpolées $P(I_i^j)$ calculées à chaque position, dans la liste d'hypothèses, d'une des n_i réponses correctes I_i^j pour la recette R_i, est alors une mesure pertinente, et pour l'ensemble des recettes la moyenne de cette précision moyenne (MAP, voir formule 6).

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{r_i} \qquad MAP = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i}P(I_i^j) \tag{6}$$

© ATALA

On rappelle que si une bonne réponse n'est pas présente dans la liste fournie par un système, sa précision moyenne vaut zéro, comme si elle avait été classée à l'infini par le système. MRR et MAP ont été calculés à l'aide du programme `trec_eval`[6].

Pour la tâche 3, rappel, précision et F-mesure ont également été calculés en prenant comme classes les titres des recettes et en examinant le titre proposé au rang 1.

3.2 Résultats des participants

Tâche 1 — Niveau de difficulté. Nous indiquons dans le tableau 3 les résultats globaux obtenus par les participants sur la première tâche. Le tableau 4 donne les résultats par classe. Les meilleurs résultats et le vainqueur sont indiqués en gras. *rf* et *nb* sont présentés en section 3.3.

Équipe, soumission	Macro				**Micro**		Rang
	R	P	F	EDRM	P=R=F	**EDRM**	
02 — LIM&Bio, #1	0,399	0,332	0,363	0,609	0,580	0,826	
02 — LIM&Bio, #2	0,415	0,339	0,373	0,624	0,586	0,828	4
02 — LIM&Bio, #3	0,406	0,329	0,364	0,621	0,574	0,824	
04 — GREYC, #1	0,273	0,250	0,261	0,501	0,489	0,794	6
04 — GREYC, #2	0,265	0,248	0,256	0,501	0,465	0,787	
04 — GREYC, #3	0,289	0,281	0,285	0,543	0,360	0,624	
22 — LIA, #1	0,353	0,633	0,453	0,600	0,592	0,828	3
22 — LIA, #2	0,363	0,603	0,453	0,643	0,581	0,813	
22 — LIA, #3	0,364	0,613	0,457	0,638	0,588	0,820	
25 — Celi France, #1	0,304	0,342	0,322	0,558	0,511	0,799	5
26 — Orange Labs, #1	0,252	0,275	0,263	0,503	0,086	0,234	
26 — Orange Labs, #2	0,395	0,524	0,451	0,659	0,612	0,840	2
28 — Wikimeta Lab, #1	**0,419**	0,460	0,438	**0,690**	0,609	0,835	
28 — Wikimeta Lab, #2	0,375	**0,682**	**0,484**	0,612	**0,625**	**0,843**	1
comparaison : rf	0,301	0,337	0,318	0,511	0,543	0,810	
comparaison : nb	0,489	0,373	0,423	0,723	0,525	0,754	

TABLE 3 – Résultats globaux sur la tâche 1, *Niveau de difficulté*, évalués en termes de rappel, précision, F-mesure et EDRM en macro-moyenne, P=R=F et EDRM en micro-moyenne

Tâche 2 — Type de plat. Le tableau 5 donne les résultats globaux tandis que le tableau 6 fournit les résultats par classe. Les meilleurs résultats et le vainqueur sont indiqués en gras.

Tâche 3 — Appariement titre/recette. Le tableau 7 donne les résultats globaux. Les meilleurs résultats et le vainqueur sont indiqués en gras. Les équipes 02 et 26 ont renvoyé plusieurs fois le même titre au rang 1 pour des recettes différentes : ces titres (= classes) sont donc proposés plusieurs fois, dont toutes sauf au mieux une sont incorrectes, ce qui fait baisser leur précision.

6. `http://trec.nist.gov/trec_eval/` : pour la moyenne de l'inverse du rang,
`trec_eval -c -q -mrecip_rank reference.qrels hypotheses.txt`,
et pour la précision moyenne non interpolée,
`trec_eval -c -q -mmap reference.qrels hypotheses.txt`.

© ATALA

Équipe	Très facile			Facile			Moyennement diff.			Difficile		
	R	P	F	R	P	F	R	P	F	R	P	F
02 #1	0,769	0,625	0,689	0,478	0,528	0,502	0,000	0,000	0,000	0,350	0,175	0,233
02 #2	0,769	0,630	0,692	0,491	0,536	0,512	0,000	0,000	0,000	0,400	0,190	**0,258**
02 #3	0,797	0,606	0,689	0,429	0,535	0,476	0,000	0,000	0,000	0,400	0,174	0,242
04 #1	0,444	0,555	0,493	0,646	0,446	0,527	0,000	0,000	0,000	0,000	0,000	0,000
04 #2	0,292	0,557	0,384	0,768	0,433	0,554	0,000	0,000	0,000	0,000	0,000	0,000
04 #3	0,232	0,575	0,331	0,506	0,440	0,471	0,418	0,107	0,170	0,000	0,000	0,000
22 #1	0,762	0,624	0,686	0,499	0,555	0,525	0,101	0,352	0,156	0,050	1,000	0,095
22 #2	0,711	0,645	0,676	0,521	0,554	0,537	0,169	0,215	0,189	0,050	1,000	0,095
22 #3	0,719	0,644	0,679	0,529	0,553	0,541	0,159	0,254	0,195	0,050	1,000	0,095
25 #1	0,498	0,586	0,539	0,617	0,465	0,530	0,101	0,317	0,153	0,000	0,000	0,000
26 #1	0,002	0,667	0,004	0,007	0,350	0,014	1,000	0,083	0,153	0,000	0,000	0,000
26 #2	0,715	0,671	0,692	0,587	0,554	**0,570**	0,180	0,472	**0,261**	0,100	0,400	0,160
28 #1	0,759	0,662	0,707	0,525	0,570	0,547	0,190	0,343	0,245	0,200	0,267	0,229
28 #2	0,786	0,660	**0,717**	0,546	0,579	0,562	0,116	0,489	0,188	0,050	1,000	0,095
rf	0,648	0,591	0,618	0,533	0,491	0,511	0,021	0,267	0,039	0,000	0,000	0,000
nb	0,640	0,651	0,646	0,427	0,560	0,484	0,339	0,202	0,253	0,550	0,077	0,136

TABLE 4 – Résultats détaillés par classe sur la tâche 1, *Niveau de difficulté*

Équipe, soumission	Macro			Micro	Rang
	R	P	F		
02 — LIM&Bio #1	0,806	0,832	0,819	0,834	
02 — LIM&Bio #2	0,841	0,842	0,841	0,849	3
02 — LIM&Bio #3	0,806	0,832	0,819	0,834	
04 — GREYC #1	0,760	0,741	0,750	0,746	5
04 — GREYC #2	0,584	0,609	0,596	0,573	
04 — GREYC #3	0,381	0,418	0,398	0,331	
22 — LIA #1	0,873	0,868	0,871	0,876	
22 — LIA #2	0,879	0,880	0,879	0,886	
22 — LIA #3	**0,881**	**0,884**	**0,882**	**0,889**	1
26 — Orange Labs #1	0,767	0,857	0,810	0,821	
26 — Orange Labs #2	0,784	0,840	0,811	0,827	4
26 — Orange Labs #3	0,779	0,789	0,784	0,784	
28 — Wikimeta Lab, #1	0,843	0,850	0,847	0,856	2
comparaison : nb	0,862	0,852	0,857	0,859	

TABLE 5 – Résultats globaux sur la tâche 2, *Type de plat*, évalués en termes de rappel, précision et F-mesure en macro-moyenne, micro-moyenne

Pour ce qui est du rappel, il est à 1 pour un titre (= une classe) si ce titre est trouvé au moins une fois correctement au rang 1 : donc c'est la proportion des titres qui sont trouvés correctement pour au moins (= exactement) une recette. Cela explique pourquoi il est égal à la précision au rang 1 (P@1), proportion des recettes dont le titre correct a été trouvé au rang 1.

© ATALA

Équipe	Entrée			Plat principal			Dessert		
	R	P	F	R	P	F	R	P	F
02 #1	0,544	0,730	0,624	0,894	0,794	0,841	0,980	0,970	0,975
02 #2	0,687	0,702	0,694	0,851	0,844	0,848	0,983	0,980	*0,982*
02 #3	0,544	0,729	0,623	0,898	0,796	0,844	0,976	0,971	0,974
04 #1	0,730	0,568	0,639	0,677	0,828	0,745	0,873	0,825	0,849
04 #2	0,662	0,378	0,481	0,540	0,699	0,609	0,551	0,749	0,635
04 #3	0,792	0,265	0,397	0,208	0,528	0,298	0,142	0,461	0,217
22 #1	0,779	0,744	0,761	0,868	0,890	0,879	0,971	0,971	0,971
22 #2	0,756	0,777	0,766	0,890	0,882	0,886	0,989	0,982	**0,986**
22 #3	0,762	0,784	**0,773**	0,893	0,883	**0,888**	0,989	0,983	**0,986**
26 #1	0,345	0,882	0,496	0,965	0,755	0,847	0,991	0,934	0,962
26 #2	0,427	0,784	0,553	0,937	0,764	0,842	0,986	0,972	*0,979*
26 #3	0,514	0,623	0,563	0,836	0,774	0,804	0,988	0,970	*0,979*
28 #1	0,669	0,742	0,703	0,872	0,840	0,856	0,989	0,969	*0,979*
nb	0,776	0,696	0,734	0,825	0,883	0,853	0,986	0,976	0,981

TABLE 6 – Résultats détaillés par classe sur la tâche 2, *Type de plat*

Équipe, soumission	Macro			P@1	MRR	Rang
	R	P	F			
02 — LIM&Bio #1	0,036	0,019	0,025	0,036	0,036	
02 — LIM&Bio #2	0,241	0,200	0,218	0,241	0,241	2
02 — LIM&Bio #3	0,133	0,119	0,125	0,133	0,133	
04 — GREYC #1	0,127	0,127	0,127	0,127	0,127	3
26 — Orange Labs #1	0,309	0,249	0,276	0,308	0,430	
26 — Orange Labs #2	0,306	0,251	0,276	0,306	0,423	
26 — Orange Labs #3	**0,314**	**0,259**	**0,284**	**0,314**	**0,434**	1

TABLE 7 – Résultats globaux sur la tâche 3, *Appariement titre-recette*, évalués en termes de rappel, précision et F-mesure en macro-moyenne, précision au rang 1, et moyenne de l'inverse du rang (MRR)

Tâche 4 — Ingrédients d'une recette. Les résultats de la dernière tâche sont donnés dans le tableau 8. Notons que les systèmes 02 et 25 n'avaient pas normalisé les noms des ingrédients qu'ils produisaient : nous avons inclus cette normalisation (désaccentuation, remplacement des espaces par des tirets), ce qui n'a pas changé le classement. La meilleure MAP de l'équipe 25 serait sinon de 0,6559 (−0,01), et celle de l'équipe 02 serait de 0,4083 (−0,05). Les équipes 25 et 28 ont également laissé passer des ingrédients hors liste normalisée (respectivement 94 et 7 dans leur meilleur système), seules les équipes 04 et 22 se sont assurées que leur système produisait uniquement des ingrédients normalisés.

© ATALA

Équipe	02 LIM&Bio	04 GREYC	22 LIA	25 Celi France	28 Wikimeta Lab
#1	0,4115	0,4881	*0,6287*	**0,6662**	0,5675
#2	0,4170	0,5074	0,6218	—	*0,6428*
#3	0,4649	0,5556	0,6191	—	—
Rang	5	4	3	1	2

TABLE 8 – Résultats globaux sur la tâche 4, *Ingrédients d'une recette*, évalués en termes de moyenne de la précision non interpolée (MAP)

3.3 Discussion

Tâche 1 — Niveau de difficulté. Les classes de cette tâche étaient très déséquilibrées : la troisième classe (Moyennement difficile) avait un nombre d'instances un ordre de grandeur au-dessous des deux premières, et la quatrième (Difficile) était encore un ordre de grandeur plus bas. De ce fait, il était beaucoup plus difficile de détecter correctement les deux dernières classes, ce que reflètent les résultats des systèmes par classe : F-mesure généralement vers 0,6–0,7 pour Très facile, 0,5 pour Facile, 0,1–0,2 pour Moyennement difficile, et 0–0,2 pour Difficile. Une bonne performance sur les deux premières classes assurait un résultat honorable en micro-mesure, et rares sont les systèmes qui ont pu détecter un nombre non négligeable d'instances des deux classes difficiles (26-#2 et #28-1).

À titre de base de comparaison, nous avons mis au point une méthode simple fondée sur une représentation des recettes par les attributs suivants. Les trois champs textuels (titre, ingrédients, préparation) ont été segmentés en mots (en conservant les nombres et ponctuations), les mots des titres ont été représentés par des attributs binaires (présence / absence du mot) et les mots des deux autres champs par des attributs numériques (tf.idf, normalisé par la longueur du champ). Le nombre d'ingrédients (nombre de <p> dans la liste des ingrédients) a également été ajouté, ainsi que la longueur en mots et caractères de la liste d'ingrédients et de la préparation. Un classifieur bayésien naïf (*nb*, NaiveBayesMultinomial dans Weka) et un classifieur à forêt d'arbres (*rf*, RandomForest dans Weka) ont été entraînés sur ces vecteurs d'attributs : ces deux classifieurs ont un temps d'entraînement court à relativement court et sont assez robustes. La forêt d'arbres a obtenu une meilleure proportion d'instances bien classées (0,544) en validation croisée sur dix parties sur le corpus d'entraînement, mais sans savoir prédire la classe Difficile ; le classifieur bayésien, avec une correction plus faible (0,527), y trouvait en revanche près de la moitié des recettes difficiles. Les résultats sur le jeu de test (bas des tableaux 3 et 4, *rf* et *nb*) reflètent ces différences, qui induisent des différences importantes dans la macro F-mesure.

Tâche 2 — Type de plat. Quatre équipes sur cinq (Bost *et al.*, 2013; Hamon *et al.*, 2013; Collin *et al.*, 2013; Charton *et al.*, 2013) ont obtenu des résultats proches de la perfection pour la catégorisation du dessert (F=0,986, 0,982, 0,979, 0,979). Cela semble indiquer que cette catégorie était plus facile à trouver, ou mieux définie : les desserts contiennent en général du sucre, etc. Pour vérifier cela, nous avons entraîné pour cette tâche le classifieur bayésien naïf avec les mêmes attributs que ci-dessus (la forêt d'arbres donnait des résultats beaucoup moins bons sur le jeu d'entraînement). Ce classifieur obtient les résultats résumés au bas des tableaux 5 et 6. Dans la catégorie Dessert, il se placerait lui aussi dans le groupe des très bons résultats. Dans les deux autres catégories, il est plusieurs points derrière le meilleur système, ce qui montre l'intérêt

 © ATALA

des traitements supplémentaires effectués par le LIA au-delà d'une simple segmentation en mots.

La distinction entre entrée et plat principal était plus difficile à établir : de fait, certains plats peuvent se servir indifféremment en entrée ou en plat principal, et les indications de quantité ne suffisent pas non plus à distinguer leur destination. Ceci étant posé, la bonne performance de cette méthode de base montre que cette tâche était nettement plus facile que la tâche 1, parce qu'elle avait une classe de moins, que ses classes étaient plus équilibrées, et que leur définition était moins subjective.

Tâche 3 — Appariement titre/recette. Les scores (MRR) pour la tâche 3 vont de 0,13 à 0,43. L'équipe 02 (Hamon *et al.*, 2013) a fourni zéro ou un titre par recette, l'équipe 04 (Lejeune *et al.*, 2013) exactement un titre par recette, et l'équipe 26 (Collin *et al.*, 2013) a produit 50 titres par recette. Cette dernière se détache des deux autres participants. Un score de 0,43 correspond à un placement du bon titre en moyenne entre les positions 2 et 3, ou encore, pour un système qui ne renverrait qu'un titre par recette, à trouver le bon titre une fois sur 2 à 3. En pratique, les systèmes des participants ont placé le bon titre en première position (P@1, précision au rang 1) dans respectivement 31,4 % des cas (éq. 26), 24,1 % des cas (02), et 12,7 % des cas (04). On vérifie ainsi que seul un système qui classe plusieurs titres pour chaque recette, comme le 26, peut obtenir un MRR supérieur à sa précision au rang 1.

Tâche 4 — Ingrédients d'une recette. Les scores (MAP) pour la tâche 4 s'étagent de 0,36 à 0,66. Le groupe des trois équipes de tête a bien réussi la tâche, avec une meilleure MAP entre 0,63 et 0,66. Ce score correspond par exemple au fait de classer en premier, dans chaque recette, les deux tiers des ingrédients, ce qui est très bon étant donné les défauts de la référence signalés plus haut.

Le tableau 9 montre le nombre de recettes pour lesquelles tous les ingrédients ont été trouvés et classés en premier (précision moyenne = 1) : les équipes 28 (Charton *et al.*, 2013) et 22 (Bost *et al.*, 2013) font beaucoup plus souvent le grand chelem que l'équipe 25 (Dini *et al.*, 2013), qui obtient pourtant la meilleure MAP (et met donc sans doute plus régulièrement les bons ingrédients vers le haut, même si ce n'est pas tout en haut). On reconnaît donc ici une stratégie

Équipe	02 LIM&Bio	04 GREYC	22 LIA	25 Celi France	28 Wikimeta Lab
#1	9	19	**173**	84	53
#2	11	68	147	—	**176**
#3	24	44	141	—	—

TABLE 9 – Tâche 4 : nombre de résultats parfaits (tous les ingrédients corrects en tête de liste)

différente entre l'équipe 25 et les équipes 22 et 28. L'équipe 25 est la seule à avoir appliqué le précepte de la MAP : classer tous les ingrédients est toujours mieux que de n'en classer qu'une partie. Elle a ainsi systématiquement classé une longue liste de quelque 850 ingrédients pour chaque recette. Si par exemple elle n'avait classé que 20 ingrédients par recette, sa MAP aurait été de 0,6513 (−0,015) — on constate que cela n'aurait néanmoins pas modifié le classement.

© ATALA

4 Conclusion

L'édition DEFT2013 a porté sur l'analyse des recettes de cuisine au travers de quatre tâches. La référence a été constituée en reprenant directement les informations présentes sur Marmiton, le site utilisé pour constituer les corpus. Du fait de cette référence « naturelle », les tâches présentaient des difficultés importantes : définitions subjectives et classes très déséquilibrées pour la difficulté de la recette, similarité entre entrée et plat principal dans les types de plats, créativité des titres de recettes, référence imparfaite et difficultés de normalisation dans la détection des ingrédients.

Au regard de ces difficultés, il apparaît que les participants ont réussi fort honorablement à prédire les valeurs renseignées par les humains qui déposent une recette sur le site. On peut même supposer que les méthodes développées pour estimer la difficulté d'une recette ou détecter ses ingrédients pourraient être utiles pour indexer automatiquement les recettes du site et lui apporter une plus grande cohérence.

Remerciements. Nous remercions Flora Badin (INIST–CNRS) et Dominic Forest (EBSI–Université de Montréal) pour les réflexions qu'ils ont menées sur les tâches à proposer aux participants du défi cette année. Nous remercions également les étudiants français (INaLCO) et canadiens (EBSI) pour les tests humains qu'ils ont réalisés et l'aide qu'ils nous ont ainsi apportée. Merci à tous les participants pour les efforts déployés et la richesse des méthodes utilisées. Nous les félicitons également pour la manière dont ils ont filé la métaphore culinaire dans leurs articles ! Enfin, nous remercions les organisateurs de TALN/Recital 2013 pour leur aide logistique dans la préparation de l'atelier de clôture de DEFT.

Références

Bost, X., Brunetti, I., Cabrera-Diego, L. A., Cossu, J.-V., Linhares, A., Morchid, M., Torres-Moreno, J.-M., El Bèze, M. et Dufour, R. (2013). Système du LIA à DEFT'13. *In Actes de DEFT*.

Charton, E., Jean-Louis, L., Meurs, M.-J. et Gagnon, M. (2013). Trois recettes d'apprentissage automatique pour un système d'extraction d'information et de classification de recettes de cuisine. *In Actes de DEFT*.

Collin, O., Guerraz, A., Hiou, Y. et Voisine, N. (2013). Participation de Orange Labs à DEFT 2013. *In Actes de DEFT*.

Dini, L., Bittar, A. et Ruhlmann, M. (2013). Approches hybrides pour l'analyse de recettes de cuisine. *In Actes de DEFT*.

Hamon, T., Périnet, A. et Grabar, N. (2013). Efficacité combinée du flou et de l'exact des recettes de cuisine. *In Actes de DEFT*.

Lejeune, G., Lecluze, C. et Brixtel, R. (2013). DEFT2013, une cuisine de caractères. *In Actes de DEFT*.

© ATALA

Efficacité combinée du flou et de l'exact des recettes de cuisine

Thierry Hamon[1] Amandine Périnet[1,2] Natalia Grabar[3]
(1) LIM&BIO (EA3969), Université Paris 13, Sorbonne Paris Cité, 74, rue Marcel Cachin, 93017 Bobigny
thierry.hamon@univ-paris13.fr
(2) Lingua et Machina c/o INRIA Rocquencourt BP 105 78153 Le Chesnay Cedex
amandine.perinet@lingua-et-machina.com
(3) CNRS UMR 8163 STL, Université Lille 1&3, 59653 Villeneuve d'Ascq
natalia.grabar@univ-lille3.fr

RÉSUMÉ

La cuisine et les recettes de cuisine occupent une place importante dans notre vie. L'alimentation a par ailleurs une influence directe sur notre bien-être, santé et bonne humeur. La compétition DEFT 2013 étant dédiée à l'analyse des recettes de cuisine, nous y avons participé pour étudier ce domaine de spécialité. Quatre tâches orientées catégorisation et extraction d'informations sont proposées. En fonction des tâches, nous utilisons des approches à base de règles et d'apprentissage. Nous avons aussi construit des ressources sémantiques pour traiter des informations exactes (par exemple, quantités exprimées en grammes ou litres, durées exprimées en minutes ou jours) et floues (par exemple, quantités exprimées en *chouilla* et *louche*, durées séquencées avec *après, ensuite, alors que*). En exploitant notre approches et nos ressources, et en naviguant entre l'exact et le flou, nous avons obtenu les résultats officiels suivants : 4e/6 sur la tâche 1, 3e/5 sur la tâche 2, 2e/3 sur la tâche 3, et 5e/5 sur la tâche 4. Plusieurs améliorations sont envisagées.

ABSTRACT

Combined efficiency of fuzziness and exactness in recipes.

An important place in our lifes is dedicated to *cuisine* and recipes. Besides, food has a direct impact on our well-being, health and mood. Because the DEFT 2013 challenge was dedicated to the analysis of recipes, we participated in this challenge to study this specialized area. Four tasks oriented on categorization and information extraction have been proposed. According to tasks, we use rule-based and machine learning approaches. We have also built semantic resources to process exact (*i.e.*, quantities expressed in grams or liters, durations expressed in minutes or days) and fuzzy (*i.e.*, quantities expressed in *chouilla* and *louche*, durations sequenced with *après, ensuite, alors que*) information. Using the approaches and resources, and sailing between the exactness and fuziness, we have obtained the following official results : 4e/6 on task 1, 3e/5 on task 2, 2e/3 on task 3, and 5e/5 on task 4. Several improvements are planned.

MOTS-CLÉS : Traitement Automatique des Langues, campagne d'évaluation, recettes de cuisine, annotation sémantique, création de ressources sémantiques, systèmes à base de règles, système d'apprentissage automatique.

KEYWORDS: Natural Language Processing, evaluation campaign, recipes, semantic annotation, semantic resource creation, rules-based system, machine learning system.

© ATALA

1 Introduction

La cuisine et les recettes de cuisine occupent une place importante dans notre vie. Elles font certainement aussi partie des facteurs principaux qui influencent notre bien-être et la bonne humeur au quotidien. En effet, selon un dicton russe, *Любовь и голод правят миром* (*L'amour et la faim gouvernent le monde*). D'autres dictons marquent aussi l'importance de la cuisine et de la nourriture dans notre vie (de Rudder, 2006; Lim, 2007) : *les carottes sont cuites*, *la fin des haricots*, *s'occuper de ses oignons* ou *avoir mangé son pain blanc*. Même nos hommes les plus illustres s'intéressent à la cuisine (Dumas, 1873).

Il n'est donc pas étonnant que les scientifiques aussi s'intéressent à l'alimentation et la nutrition. Des phénomènes connus au grand public comme ceux de *malbouffe* ou de *French paradox* sont étudiés. Si ces études expliquent et justifient les bienfaits du vin rouge et de ses substances, elles montrent aussi que la malbouffe (par exemple, la consommation de hamburgers, de hot-dogs, de frites, de chips ou de sodas) peut favoriser l'obésité, le diabète, les maladies cardiovasculaires, des dépressions et même certains cancers. Il ne faut donc pas sous-estimer la place que l'alimentation, et la bonne alimentation surtout, occupent dans notre vie. Pour mieux étudier les questions relatives à l'alimentation de la population, il existe même une sous-section 44.04 de CNU dédiée à la nutrition. L'effort principal des chercheurs consiste à attirer notre attention sur la qualité de notre alimentation et son influence sur notre santé. Depuis quelques années, l'étude NutriNet-Santé [1] semble avoir concentré les efforts et les initiatives autour de ces questions en France. Dans de telles études, une attention particulière peut être portée à certaines pathologies, comme le diabète, à la condition sociale ou à la provenance géographique de la population étudiée (Jones-Smith *et al.*, 2013; Andreeva *et al.*, 2013).

Dans les domaines de l'informatique et du Traitement Automatique de Langues (TAL), les textes de recettes de cuisine ont aussi attiré l'attention des chercheurs. Le premier travail est certainement celui décrivant le système Epicure et son application aux recettes de cuisine (Dale, 1989). Ce système a pour objectif de générer une description linguistique des recettes de cuisine. Des représentations syntaxiques profonde et de surface sont proposées en utilisant la grammaire d'unification. Le fonctionnement du système s'appuie sur des connaissances du domaine, comme par exemple les objets (individuels, massifs, quantifiés ou non) encodés dans une ontologie. Les objets évoluent en fonction des états qu'ils subissent. Ce système de TAL prend aussi en compte la pronominalisation et les anaphores. L'objectif de ce système semble d'avoir implémenté une approche sémantique fine. Le passage à l'échelle et le traitement de grosses données n'étaient pas au centre d'intérêt à l'époque.

Un travail plus récent est fait dans le cadre du Computer Cooking Contest [2], qui a lieu tous les ans depuis 2008. Cette compétition propose de traiter les recettes de cuisine. En 2012, les objectifs étaient clairement orientés sur le Raisonnement à partir de cas. Par exemple, il s'agit de construire automatiquement un système à base de cas (1) d'une part pour en montrer la faisabilité à partir du texte libre et (2) d'autre part pour proposer de nouvelles recettes en se basant sur les recettes existantes et en fonction des spécifications données (Dufour-Lussier *et al.*, 2013). Dans ce travail cité, l'attention particulière est réservée aux actions (souvent des verbes) et à leurs arguments. En plus, plusieurs traitements de TAL sont appliqués : étiquetage morpho-syntaxique, résolution d'anaphores, analyse syntaxique. 15 recettes de cuisine sont traitées. Cette compétition

1. *https ://www.etude-nutrinet-sante.fr*
2. *http ://computercookingcontest.net*

© ATALA

a donné aussi lieu aux travaux interdisciplinaires sur la reconnaissance d'aliments dans notre frigo (Kamoda *et al.*, 2012), ou encore sur l'apprentissage et modélisation des gestes culinaires grâce à des gants spéciaux (Ota *et al.*, 2012).

La compétition de TAL DEFT 2013 propose aussi de travailler avec les recettes de cuisine. Devant cette perspective de découverte et de mise au point scientifique, culturelle et culinaire, nous avons décidé de participer au challenge. Les quatre tâches sympathiques de la compétition sont orientées sur la catégorisation des recettes (tâches 1 et 2) et l'extraction d'information (tâches 3 et 4). Ces tâches sont les suivantes :

1. Identifier à partir du titre et du texte de la recette son niveau de difficulté sur une échelle à 4 niveaux : très facile, facile, moyennement difficile, difficile.
2. Identifier à partir du titre et du texte de la recette le type de plat préparé : entrée, plat principal, dessert.
3. Apparier le texte d'une recette à son titre.
4. Extraire du titre et du texte d'une recette la liste de ses ingrédients.

Il va de soi que, pour aborder efficacement et positivement le traitement des recettes de cuisine, il faut être muni de bons ingrédients linguistiques (section 2) et de bons ustensiles de TAL (section 3). Nous présentons aussi les recettes proposées (section 4) et les résultats que nous avons pu obtenir en sortie (section 5). Comme le perfectionnement (au moins scientifique, culturel et culinaire) n'a pas de limites, nous mentionnons quelques perspectives que nous aimerions mettre en oeuvre dans l'avenir proche (section 6).

2 Ingrédients

Parmi les ingrédients utilisés, nous avons plusieurs ressources (résumées dans la table 1) :

– Une liste d'ingrédients de cuisine et d'aliments collectés dans des sources disponibles en ligne [3 4 5 6], dans la partie française de l'UMLS (NLM, 2011) et à partir des ingrédients connus dans l'ensemble d'entraînement de la campagne. Une des difficultés a été de distinguer entre les ingrédients et les aliments. La solution que nous avons adoptée est de considérer que les ingrédients sont "bruts" alors que les aliments sont déjà préparés voire cuisinés. Logiquement, ce sont les ingrédients qui sont utilisés dans les recettes... Ce qui n'est bien sûr pas totalement vrai car de bons aliments comme les pâtes, les raviolis, les glaces, les sauces diverses et variées, sans parler des fromages, font aussi de très bons ingrédients irremplaçables dans notre cuisine. Devant ce flou inattendu, nous avons finalement décidé (1) d'introduire l'ambiguïté et de catégoriser les aliments concernés comme ingrédients aussi, ou bien (2) de simplifier la réalité et de les considérer comme de vrais ingrédients. Cette liste comporte 6 070 entrées catégorisées en viandes, légumes, fruits, produits de boulangerie, poissons, sucreries, etc.
– Une liste d'ustensiles collectés à partir de ressources disponibles en ligne [7 8]. Cette liste comporte 222 entrées.

3. *http ://www.bioweight.com/glucides.html*
4. *http ://www.bioweight.com/proteines.html*
5. *http ://www.centre-clauderer.com/acides-bases/femme-2.htm*
6. *http ://les.calories.free.fr/*
7. *http ://popoblog.unblog.fr/liste-ustensiles-de-cuisine-mise-a-jour-le-130808/*
8. *fr.wikipedia.org/wiki/Ustensile_de_cuisine*

© ATALA

– Une liste de mots vides du français composée de 616 entrées et une liste d'exceptions composée de 37 entrées.
– Une ressource pour la détection de quantités des ingrédients, comme par exemple : *250 g de sucre, 3 oeufs, une bonne cuillère d'huile, 100 + 50 gr de beurre, 1/2 l de lait, deux graines de cardamone, un chouilla de sel, 2 louches de bouillon, beaucoup de menthe*. Nous avons distingué les quantités standards, déjà exprimées en grammes ou litres, et les quantités non standards (Delamasure, 2007), non exprimées en grammes ou en litres, mais qui font le charme de nos recettes de cuisine et en garantissent la réussite. Afin d'en assurer un traitement égalitaire et une utilisation possible dans un système de TAL, nous avons établi des heuristiques et avons converti toutes ces quantités en litres ou en kilogrammes. Nous nous sommes aidés pour ceci d'autres ressources et convertisseurs disponibles en ligne [9] [10]. Lorsque plusieurs équivalences étaient connues ou lorsque la question de conversion ne s'était pas encore posée, nous avons avancé nos propres propositions. Par exemple, les expressions comme *pincée, soupçon, chouilla, lichette, rasade, giclée, goutte, microgramme, 1 peu* signifient que la recette comporte 1 gramme de produit ; les expressions comme *louche, tube, fond, ravier, assiette creuse, poignée* signifient que la recette comporte 100 grammes de produit, etc. Lors de la normalisation, plusieurs opérations arithmétiques sont appliquées (somme, division, multiplication).
– Une ressource pour détecter des durées. Là aussi, nous distinguons la durée quantifiée et exprimée en minutes, heures, jours, nuits, etc., et la durée non quantifiée et séquencée par des expressions comme *puis, pendant que, alors que, ensuite*.
– Une liste d'actions (verbes et noms) de la ressource Verbaction (Hathout *et al.*, 2001).
– Une ressource distributionnelle (Harris, 1954) construite à partir des titres et du contenu des recettes (ensembles d'entraînement et de test). L'approche s'appuie sur deux méthodes d'acquisition de relations d'hyperonymie (Morin et Jacquemin, 2004; Bodenreider *et al.*, 2001) et une méthode l'acquisition de variantes terminologiques (Jacquemin, 1996). Pour délimiter les contextes de mots, nous utilisons des fenêtres graphiques de 10 mots à droite et 10 mots à gauche, en limitant les mots en relation aux mots appartenant à la même catégorie syntaxique (noms, adjectifs et termes identifiés par l'extracteur YaTeA (Aubin et Hamon, 2006)). En ce qui concerne la mesure de similarité, nous utilisons l'indice de Jaccard (Grefenstette et Tapanainen, 1994), qui normalise le nombre de contextes partagés par deux mots par le nombre total de contextes de ces mots. Nous avons également filtré les relations acquises avec la moyenne de trois seuils (nombre de contextes partagés, fréquence des contextes partagés, fréquence des mots pivots).
– Une ressource distributionnelle *FreDist* (Anguiano et Denis, 2011).
– Une ressource de synonymes de la langue générale fournis par le Petit Robert (Robert, 1990).

L'ingrédient principal reste bien sûr l'ensemble d'entraînement fourni par les organisateurs. Cet ensemble comporte 13 864 recettes. Chaque recette contient les informations suivantes : le titre, les ingrédients, les étapes ou le déroulement de la préparation, le coût, le niveau de difficulté et le type de plat. Mais seule le coût est disponible dans l'ensemble de test. Souvent, la boisson conseillée est aussi indiquée. L'ensemble d'entraînement, aidé par les ressources, a permis de créer le système de TAL et d'en faire les tests. Un autre ingrédient important sont les recettes du corpus de test, au nombre de 9 200 (2 300 par tâche). Ce corpus a été fourni pour une durée de trois jours : comme le feu de nos cuisinières ne chauffait pas assez fort et vite et pour ne pas

9. *http ://www.supertoinette.com/mesures-equivalences-culinaires.html*
10. *http ://webcafe.highbb.com/t1827-mesures-metriques-et-nord-americaine#13245*

© ATALA

Ressources	Exactes	Floues
Ressources lexicales	+ (ingrédients, ustensiles, synonymes)	+ (distributionnelles, aliments)
Quantités	+ exprimées en grammes et litres	+ exprimées autrement
Durées	+ quantifiées	+ non quantifiées
Actions	+	
Mots vides	+	
Recettes de cuisine	+ ensemble d'entraînement	+ ensemble de test

TABLE 1 – Ressources construites et utilisées afin de pouvoir traiter les informations exactes et floues contenues dans les recettes de cuisine.

risquer l'indigestion, le système de TAL était le seul moyen de test possible pour traiter les quatre tâches.

3 Ustensiles de TAL

Nos ustensiles de TAL sont composés d'un système à base de règles et d'un système d'apprentissage automatique.

3.1 Système à base de règles

L'objectif du système à base de règles est d'assurer la reconnaissance de mots clés (ingrédients, ustensiles de cuisine...) et d'informations associées (quantités d'ingrédients, durées de préparation...). Ce système a quatre étapes principales. Trois étapes (reconnaissance de termes et d'informations associées, pondération de termes, et leur filtrage et sélection) sont adaptées du système utilisé lors de la compétition DEFT 2012 (Hamon, 2012). La quatrième étape, dédiée à l'appariement de titres, est adaptée du travail précédent (Hamon et Gagnayre, 2012).

3.1.1 Reconnaissance de termes (TermTagger) et d'informations associées

Les ressources décrites plus haut (listes d'ingrédients, d'ustensiles et d'actions) sont projetées sur les textes traités. Cette projection prend en compte les lemmes de mots du corpus. Nous utilisons le module Perl `Alvis::TermTagger` [11]. Parallèlement à la reconnaissance des termes, les entités nommées associées sont aussi recherchées. Par exemple, nous exploitons les ressources décrites plus haut pour reconnaître les quantités des ingrédients ou bien les durées nécessaires à la préparation d'un recette.

11. `http://search.cpan.org/~thhamon/Alvis-TermTagger/`

© ATALA

3.1.2 Pondération de termes

Pour identifier les termes les plus importants parmi ceux qui sont reconnus, nous les trions par ordre de pertinence avec plusieurs méthodes de pondération :

- La fréquence du terme dans le document (**tf**) ;
- Le nombre de documents du corpus où le terme apparaît (**df**) ;
- La position de la première occurrence du terme (**position**). Nous considérons ici que les termes situés au début du document ont un poids plus élevé que ceux situés à la fin ;
- Le cosinus de la position de la première occurrence du terme (**positionCos**). Nous faisons l'hypothèse que les termes qui apparaissent au début ou à la fin du texte sont les plus pertinents. Nous plaçons ainsi la première occurrence de chaque terme sur le cercle trigonométrique en considérant que le début du document est l'angle 0 et la fin l'angle 2π. Nous avons calculé le cosinus de l'angle formé par la position ;
- La fréquence de la forme canonique (lemmatisée) du terme (**canon**) ;
- Le fait que les termes sont quantifiés (**quant**). Ceci est essentiellement le cas des ingrédients.

3.1.3 Filtrage et sélection des termes

Les termes peuvent ensuite être filtrés ou regroupés suivant différents critères :

- Suppression des termes isolés étiquetés comme adjectifs (**filtrAdj**). Il s'agit alors de modifieurs de termes complexes, qui n'ont pas lieu d'apparaître tout seuls. Les adjectifs correspondant à des termes des ressources sont conservés ;
- Prise en compte de l'inclusion lexicale (**filtrInclLex**). Les termes en position tête d'un terme et ayant de rang plus élevé dans la liste triée sont supprimés. Par exemple, nous traitons de cette manière les termes comme {*crème anglaise*, *crème*}, {*pomme de terre*, *pomme*}, {*fève de tonka*, *fève*}. Les calculs des inclusions sont effectués au niveau syntaxique et au niveau des chaînes de caractères, lorsque la syntaxe ne détecte pas de relations ni d'inclusions ;
- Regroupement des termes en fonction de leur forme canonique (**filtrCan**). Il s'agit du regroupement des lemmes des composants et du filtrage par la forme fléchie la plus fréquente.

3.1.4 Appariement de titres

L'appariement de titres avec les recettes est effectué suivant les étapes suivantes :

- *Pré-traitement des titres et leur conversion en mots-clés.* À cette étape, les titres sont segmentés en mots, étiquetés morpho-syntaxiquement et lemmatisés (Schmid, 1994). En fonction des expériences, tous les mots ou seulement les mots "significatifs" (noms, adjectifs et verbes) sont retenus. Dans ce dernier cas, les mots grammaticaux et les mots de la liste d'exceptions ne sont pas considérés. Par exemple, le titre *Tarte au caramel et aux bananes* est converti en mots-clés suivants (*tarte, caramel, banane*).
- *Enrichissement de mots-clés avec les ressources linguistiques.* Les mots-clés sont étendus avec les ressources linguistiques : synonymes ou les mots associés selon les ressources. Chaque mot-clé, et les mots faisant partie de son extension, forment un cluster. Par exemple, le mot-clé *tarte* est enrichi avec *gâteau, pâtisserie, tartelette, gaufre, cake, galette, crêpe, beignet*, le mot-clé *banane* avec *pomme, fruit, abricot, poire*, et le mot-clé *caramel* avec *chocolat, beurre, vanille, croûton,*

 © ATALA

pâte, sirop. Les mots-clés sont considérés comme les labels sémantiques de leurs clusters : *tarte, banane, caramel*, respectivement. De cette manière, chaque titre est décrit par n clusters, en fonction du nombre de ses mots-clés.

– *Pré-traitement des recettes et leur sélection.* Les recettes sont pré-traitées de la même manière que les titres, et segmentées en phrases. Ensuite, les mots et termes des clusters sont appariés avec les mots et les termes des recettes. C'est l'étape la plus importante : en plus de l'expansion et de l'appariement, une sélection des appariements entre les titres et les recettes est effectuée. Notons qu'à cette étape, il est possible d'effectuer l'appariement complet ou partiel (avec un pourcentage donné) entre les titres et les recettes. Par exemple, nous pouvons apparier le titre *Tarte au caramel et aux bananes* à la recette 90954 grâce à l'appariement direct et partiel (le mot-clé *tarte* n'apparaît pas dans le texte de la recette).

3.2 Système d'apprentissage automatique

Le système d'apprentissage est basé sur les fonctionnalités proposées par la plate-forme Weka (Witten et Frank, 2005). Plus particulièrement, nous utilisons et effectuons :

– le sur-échantillonnage avec l'algorithme *SMOTE* (Chawla *et al.*, 2002), particulièrement utile sur la tâche 1 (difficulté des recettes), où il existe un déséquilibre net dans les données ;
– la sélection d'attributs (avec *CFS combiné avec BestFirst* (Hall, 1998)), particulièrement utile pour rendre possibles les traitements étant donné le nombre total d'attributs ($n > 20\ 000$) ;
– les algorithmes d'apprentissage testés sont : SVM, arbres de décision, régression logistique ;
– les attributs exploités sont assez variés, par exemple :
  – les lemmes ($n = 8\ 145$) ;
  – les formes fléchies qui ne sont pas les lemmes ($n = 7\ 830$) ;
  – les étiquettes morpho-syntaxiques ($n = 34$) ;
  – les étiquettes sémantiques provenant des ressources ($n = 61$), dont les types d'ingrédients, les ustensiles, les verbes d'action ;
  – le coût estimé d'une recette ;
  – la taille de la recette en nombre de caractères et de mots ;
  – le nombre d'étapes explicites et implicites de la recette ;
  – le nombre total d'étapes dans une recette ;
  – le nombre de lignes dans les ingrédients lorsque disponibles ;
  – la liste des actions ;
  – le nombre de documents du corpus où le terme apparaît (**df**) ;
  – le nombre d'occurrences des notions suivantes : sucre, sel, poivre, piment, huile, moutarde. Par exemple, lorsque *sucre, cassonade, sucre roux* sont reconnus, ils incrémentent la notion de sucre ;
  – les indications de temps exactes ou floues ;
  – les quantités exactes ou floues.

4 Recettes et tests réalisés

Pour aborder les tâches de la compétition, et comme *chaque tache a sa recette*, nous avons exploité le corpus d'entraînement pour mettre au point des stratégies différentes selon les tâches.

© ATALA

4.1 Tâche 1 : niveau de difficulté

Les textes des recettes sont d'abord traités avec le système à base de règles (section 3.1). Ensuite le système d'apprentissage (section 3.2) est appliqué. Les attributs exploités sont les suivants :
– les lemmes ;
– les étiquettes morpho-syntaxiques ;
– les étiquettes sémantiques provenant des ressources ;
– le coût estimé d'une recette ;
– la taille de la recette en nombre de caractères et de mots ;
– le nombre d'étapes explicites et implicites de la recette ;
– le nombre total d'étapes dans une recette ;
– le nombre de lignes dans les ingrédients lorsque disponibles ;
– la liste des actions ;
– le nombre de documents du corpus où le terme apparaît (**df**) ;
– les indications de temps exactes ou floues ;
– les quantités exactes ou floues.

Il existe un déséquilibre dans les données du corpus d'entraînement, où les plat difficiles sont clairement sous-représentés :
– plats très faciles : 6 962,
– plats faciles : 5 752,
– plats moyennement difficiles : 1 068,
– plats difficiles : 80.

Nous effectuons donc un sur-échantillonage avec *SMOTE* : +1000% pour la classe 4, +200% pour la classe 3, combinaison des deux, tests d'autres valeurs par défaut. Nous pouvons ainsi constaté qu'il existe une amélioration nette lorsque le sur-échantillonage est effectué sur la classe 4 (la moins bien représentée).

Nous effectuons aussi une sélection d'attributs et pouvons par exemple constater que :
– très peu d'attributs apparaissent utiles ($n = 53$) : catégories morpho-syntaxiques, des lemmes, mais très peu de formes fléchies ce qui explique qu'elles ne sont pas utilisées,
– les catégories sémantiques ne sont pas saillantes,
– la taille et les sections des recettes sont utiles,
– les actions et le coût se révèlent importants aussi,
– les durées et le temps, exact ou flou, sont utiles,
– de même que la fréquence dans le corpus total de certains ingrédients.

4.2 Tâche 2 : type de plat

Les textes des recettes sont d'abord traités avec le système à base de règles (section 3.1). Ensuite le système d'apprentissage (section 3.2) est appliqué. Les attributs exploités sont les suivants :
– les lemmes ($n = 8\ 145$) ;
– les étiquettes morpho-syntaxiques ($n = 34$) ;
– les étiquettes sémantiques provenant des ressources ;
– le coût estimé d'une recette ;

© ATALA

– la taille de la recette en nombre de caractères et de mots ;
– le nombre d'étapes explicites et implicites de la recette ;
– le nombre total d'étapes dans une recette ;
– le nombre de lignes dans les ingrédients lorsque disponibles ;
– la liste des actions ;
– la fréquence des ingrédients dans le corpus total (**df**) ;
– le nombre d'occurrences des notions suivantes : sucre, sel, poivre, piment, huile, moutarde.
– les indications de temps exactes ou floues ;
– les quantités exactes ou floues.

Nous avons testé le sur-échantillonnage, mais comme les données sont équilibrées pour cette tâche (3 246 entrées, 6 449 plats, 4 169 desserts), le sur-échantillonnage ne fournit pas de gain.

Nous effectuons aussi la sélection d'attributs et pouvons constater que :
– une réduction importante du nombre d'attributs est effectuée ($n = 72$),
– beaucoup de lemmes sont gardés, mais seulement 4 formes fléchies, qui ne seront donc pas utilisées pour cette tâche non plus,
– les catégories morpho-syntaxiques ne semblent pas être saillantes,
– les catégories sémantiques ne sont pas importantes,
– les quantités (exactes ou floues) de certains ingrédients (sel, poivre, moutarde) sont utiles,
– les indications sur les durées ne sont pas saillantes,
– certaines actions (comme *entrer*), dont la nominalisation est le type de plat *entrée*,
– le nombre de documents où apparaît l'ingrédient.

4.3 Tâche 3 : appariement du texte d'une recette à son titre

L'appariement des titres avec les textes des recettes est effectué avec l'approche décrite dans la section 3.1.4. Nous testons l'influence des ressources sémantiques (synonymes, distributionnelles) et constatons par exemple que les synonymes détériorent les résultats. Nous testons aussi l'utilisation de tous les mots ou des mots "significatifs" seulement et constatons que dans ce dernier cas, les résultats sont améliorés. Comme l'appariement à 100 % ne couvre pas la moitié de recettes, nous appliquons l'appariement partiel, ce qui permet d'augmenter la couverture. Afin de limiter l'appariement de plusieurs recettes à un même titre, lorsqu'un titre est affecté à une recette, celui-ci est retiré de l'ensemble des titres à assigner. Pour une recette donnée, les titres candidats sont ordonnés en fonction du rapport entre le nombre de mots appariés et le nombre maximum de mots appariés parmi tous les titres candidats. Parmi tous les titres candidats, le choix du titre est effectué de la manière suivante :
– le premier titre candidat, qui n'a pas déjà été assigné, est sélectionné,
– lorsqu'il existe aucun titre candidat non assigné, c'est le premier titre candidat qui est sélectionné.

Pour les recettes qui n'ont pas de titres avec des correspondances lexicales appariées ou bien lorsque cet appariement est inférieur à un seuil donné (par exemple, 75 % ou 50 %), aucun titre ne leur est assigné.

© ATALA

4.4 Tâche 4 : extraction de la liste des ingrédients

Pour l'extraction des ingrédients du texte des recettes, nous exploitons les étapes 3.1.1 à 3.1.3 du système basé sur les règles. Cela veut dire que les ressources sémantiques sont projetées sur les textes des recettes afin d'y reconnaître les ingrédients. Ensuite, les ingrédients sont pondérés avec les différentes méthodes. La solution qui s'avère être la plus efficace est la suivante : **position** * (**tf** des ingrédients quantifiés). Lorsque plusieurs ingrédients ont un poids égal, ils sont ordonnés en fonction de la fréquence **canon** dans la recette.

Les difficultés qui restent avec cette tâche sont :
– hésitation entre les formes courtes et étendues des ingrédients,
– hésitation entre les formes fléchies et lemmatisées des ingrédients.

Le calcul des inclusions lexicales syntaxiques et au niveau des chaînes de caractères a été effectué à cet effet. Cela a permis de faire des tests systématiquement, mais il reste difficile de reproduire la logique des données de référence et de la section *Ingrédients* des recettes. Le plus souvent, notre approche permet d'extraire les ingrédients mais leurs formes ou positions pondérées peuvent ne pas être correctes.

5 Résultats obtenus

Nous avons soumis trois runs pour chacune des quatre tâches de la compétition. Notre classement officiel est le suivant :
– tâche 1 : 4e/6
– tâche 2 : 3e/5
– tâche 3 : 2e/3
– tâche 4 : 5e/5

Les résultats détaillés pour ces tâches sont indiqués dans le tableau 2.

Selon les organisateurs de la compétition :
– Sur la première tâche (niveau de difficulté), six équipes ont participé. Les résultats globaux obtenus en termes de micro-mesure varient de 0,625 à 0,489 sur les meilleures soumissions de chaque équipe (micro mesure moyenne = 0,569, médiane = 0,589).
– Sur la deuxième tâche (type de plat), cinq équipes ont participé. Les résultats globaux obtenus varient entre 0,889 et 0,746 (micro-mesure) sur les meilleures soumissions de chacun (moyenne de 0,833, médiane de 0,849).
– Sur la troisième tâche (appariement titre/recette), trois équipes ont participé. Les micro-mesures varient de 0,314 à 0,127 (moyenne de 0,227, médiane de 0,241). Le MRR (Mean Reciprocal Rank) calculé varie de 0,434 à 0,196 sur ces mêmes soumissions.
– Sur la quatrième tâche (normalisation des ingrédients), cinq équipes ont participé. Les résultats, évalués en termes de MAP (Mean Average Precision) varient, sur les meilleures soumissions de chaque équipe, entre 0,6662 et 0,4649 (moyenne de 0,5916, médiane de 0,6287).

© ATALA

Tâche/Run	Macro			Micro			MRR/MAP
	R	P	F	R	P	F	
T1/R1	0.399	0.332	0.363	0.580	0.580	0.580	
T1/R2	0.415	0.339	0.373	0.586	0.586	**0.586**	
T1/R3	0.406	0.329	0.364	0.574	0.574	0.574	
T2/R1	0.806	0.832	0.819	0.834	0.834	0.834	
T2/R2	0.841	0.842	0.841	0.849	0.849	**0.849**	
T2/R3	0.806	0.832	0.819	0.834	0.834	0.834	
T3/R1	0.036	0.019	0.025	0.036	0.036	0.036	0.0360
T3/R2	0.241	0.200	0.218	0.241	0.241	**0.241**	**0.2413**
T3/R3	0.133	0.119	0.125	0.133	0.133	0.133	0.1326
T4/R1							0.4115/0.4543
T4/R2							0.4170/0.4596
T4/R3							**0.4649**/0.4705

TABLE 2 – Performances officielles sur les quatre tâches sur le corpus de test

Par rapport à d'autres participants, nous sommes au-dessus des moyennes sauf pour la tâche 4. Sur la tâche 4, les résultats produits comportent un défaut de format. Les résultats obtenus après la correction sont : 0.4543, 0.4596, 0.4705 pour les trois runs respectivement. Ces nouveaux résultats sont indiqués après / dans le tableau.

5.1 Tâche 1 : niveau de difficulté

Les runs évalués ont été obtenus avec l'algorithme SVM, le sur-échantillonnage avec *SMOTE*, et la sélection d'attributs. La différence entre les runs est la suivante :
– run 1 : tous les attributs sont utilisés,
– run 2 : les actions ne sont pas utilisées,
– run 3 : les lemmes ne sont pas utilisés, mais les actions sont utilisées.

Au sein de ces tests, c'est la configuration SVM, sur-échantillonnage, sélection d'attributs, et sans les actions qui montre de meilleurs résultats.

5.2 Tâche 2 : type de plat

Les runs évalués ont été obtenus avec l'algorithme SVM. La différence entre les runs est la suivante :
– run 1 : sélection d'attributs, utilisation des actions,
– run 2 : pas de sélection d'attributs, utilisation des actions,
– run 3 : sélection d'attributs, sans l'utilisation des actions.

C'est la configuration SVM, sans sélection d'attributs, et avec utilisation des actions qui donne les meilleurs résultats.

© ATALA

5.3 Tâche 3 : appariement du texte d'une recette à son titre

Les paramètres des runs de la tâche 3 sont :
- Run 1 : catégories majeures (noms, verbes, adjectifs) ; prise en compte de la liste d'exceptions, appariement partiel à 75 % ;
- Run 2 : tous les mots du titre, prise en compte de la liste d'exceptions, appariement partiel à 50 % ;
- Run 3 : catégories majeures (noms, verbes, adjectifs) ; prise en compte de la liste d'exceptions ; utilisation de la ressource distributionnelle ; appariement complet.

C'est le run 2 qui a montré de meilleurs résultats sur les données de test.

5.4 Tâche 4 : extraction de la liste des ingrédients

Pour cette tâche, nous utilisons le module Perl TermTagger. Les paramètres communs des runs soumis sont :
- utilisation de la liste des ingrédients constituée à partir de ressources en ligne,
- assignation du poids comme spécifié dans la section 4.4,
- les termes les plus grands calculés avec les inclusions syntaxiques.

La différence entre les runs est la suivante :
- run 1 : ajout de la liste des ingrédients du corpus d'entraînement,
- run 2 : paramétrage par défaut,
- run 3 : ajout des termes les plus grands calculés avec les inclusions au niveau des chaînes de caractères, et utilisation de FLEMM (Namer, 2000).

Le meilleur run est obtenu avec l'ajout des termes les plus grands calculés avec les inclusions au niveau des chaînes de caractères, et avec l'utilisation de FLEMM. Comme nous l'avons noté plus haut, souvent notre approche permet d'extraire les ingrédients mais leurs formes ou positions pondérées peuvent ne pas être correctes. Il reste en effet difficile de reproduire la logique des données de référence.

6 Conclusions et perspectives de perfectionnement

Nous avons présenté les expériences et tests effectués dans le cadre de la compétition DEFT 2013 dédiée au traitement des recettes de cuisine. Il s'agit d'un domaine intéressant où les informations exactes et floues co-existent et garantissent certainement la réussite de nos recettes de cuisine. Face aux systèmes de TAL, ces informations doivent être traitées de manière spéciale. Nous avons ainsi effectué une normalisation du flou vers de l'exact, en nous inspirant de la norme pifométrique dédiée (Delamasure, 2007) et des instructions précieuses trouvées en ligne. L'approche développée est basée sur un système à base de règles et un système d'apprentissage, de même que différentes ressources. Nous avons obtenus les résultats officiels suivants : 4e/6 sur la tâche 1, 3e/5 sur la tâche 2, 2e/3 sur la tâche 3, et 5e/5 sur la tâche 4.

© ATALA

En travaillant sur les recettes de cuisine, nous nous sommes rendus compte que ces recettes sont bien plus que de simples instructions pour bien remplir nos petits ventres. Les recettes de cuisines deviennent en effet une vitrine d'expression. Par exemple, nous pouvons y trouver l'appel à la consommation de la viande des alligators du Canada, la fierté pour les recettes familiales proposées, les conseils de vie et de bonne nourriture, sans parler des émotions, de l'humour et des opinions.

Il reste plusieurs perspectives à notre travail. Nous avons plusieurs perspectives scientifiques :
- tester mieux l'influence de différents attributs dans les tâches 1 et 2 ;
- faire d'autres tests sur l'ordonnancement des ingrédients grâce à l'utilisation des CRF dans la tâche 4 ;
- effectuer d'autres tests avec les ressources pour l'expansion de titres et l'ordonnancement de titres dans la tâche 3 ;
- prendre en compte d'autres facteurs du flou comme les erreurs d'orthographe ;
- prendre en compte les anaphores ;
- exploiter les relations sémantiques entre les ingrédients pour réduire le nombre d'attributs.

Parmi les perspectives culinaires, nous aimerions tester autant de recettes que possible, moyennant la possibilité de trouver les ingrédients nécessaires (viandes d'alligator, jésus...).

Références

ANDREEVA, V., MARTIN, C., ISSANCHOU, S., HERCBERG, S., KESSE-GUYOT, E. et MÉJEAN, C. (2013). Sociodemographic profiles regarding bitter food consumption. cross-sectional evidence from a general french population. *Appetite*, 67(3):53–60.

ANGUIANO, E. et DENIS, P. (2011). FreDist : Automatic construction of distributional thesauri for French. *In TALN*, pages 119–124.

AUBIN, S. et HAMON, T. (2006). Improving term extraction with terminological resources. *In* SALAKOSKI, T., GINTER, F., PYYSALO, S. et PAHIKKALA, T., éditeurs : *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, numéro 4139 de LNAI, pages 380–387. Springer.

BODENREIDER, O., BURGUN, A. et RINDFLESCH, T. (2001). Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS. *In* URI INIST CNRS, éditeur : *Terminologie et Intelligence artificielle (TIA)*, pages 11–21, Nancy.

CHAWLA, N. V., BOWYER, K. W., HALL, L. O. et KEGELMEYER, W. P. (2002). Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

DALE, R. (1989). Cooking up referring expressions. *In Annual meeting on Association for Computational Linguistics*, pages 68–75.

DE RUDDER, O. (2006). *Aux petits oignons ! : Cuisine et nourriture dans les expressions de la langue française*. Larousse.

DELAMASURE, M. (2007). Norme française NF UNM 00-003. système d'unités pifométriques. Rapport technique, NA.

DUFOUR-LUSSIER, V., LE BER, F., LIEBER, J. et NAUER, E. (2013). Automatic case acquisition from texts for process-oriented case-based reasoning. *Information Systems*. Comming soon.

 © ATALA

DUMAS, A. (1873). *Le grand dictionnaire de cuisine*. Editions Pierre Grobel.

GREFENSTETTE, G. et TAPANAINEN, P. (1994). What is a word, what is a sentence ? Problems of tokenization. *In COMPLEX (Computational lexicography and text research)*, pages 79–87.

HALL, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*. Thèse de doctorat, University of Waikato, Hamilton, New Zealand.

HAMON, T. (2012). Acquisition terminologique pour identifier les mots clés d'articles scientifiques. *In Actes de l'atelier DEFT 2012*, pages 25–31, Grenoble, France.

HAMON, T. et GAGNAYRE, R. (2012). Linking expert and lay knowledge with distributional and synonymy resources : application to the mining of diabetes fora. *In Proceedings of The Fourth Swedish Language Technology Conference (SLTC 2012)*, pages 35–36, Lund, Sweden.

HARRIS, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

HATHOUT, N., NAMER, F. et DAL, G. (2001). An experimental constructional database : the MorTAL project. *In* BOUCHER, P., éditeur : *Morphology book*. Cascadilla Press, Cambridge, MA.

JACQUEMIN, C. (1996). A symbolic and surgical acquisition of terms through variation. *In* WERMTER, S., RILOFF, E. et SCHELER, G., éditeurs : *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 425–438, Springer.

JONES-SMITH, J., KARTER, A., WARTON, E., KELLY, M., KERSTEN, E., MOFFET, H., ADLER, N., SCHILLINGER, D. et LARAIA, B. (2013). Obesity and the food environment : Income and ethnicity differences among people with diabetes : The diabetes study of Northern California (DISTANCE). *Diabetes Care*, 36(1):1200–1208.

KAMODA, R., UEDA, M., FUNATOMI, T., IIYAMA, M. et MINOH, M. (2012). Grocery re-identification using load balance feature on the shelf for monitoring grocery inventory. *In Proceedings of the Cooking with Computers workshop (CwC)*, pages 8–18.

LIM, J. S. (2007). Description structurale des proverbes coréens autour des « noms de cuisine ». *In Proceedings of the 26th conference on Lexis and Grammar*, Bonifacio.

MORIN, E. et JACQUEMIN, C. (2004). Automatic Acquisition and Expansion of Hypernym Links. *Computers and the Humanities*, 38(4):363–396.

NAMER, F. (2000). FLEMM : un analyseur flexionnel du français à base de règles. *Traitement automatique des langues (TAL)*, 41(2):523–547.

NLM (2011). *UMLS Knowledge Sources Manual*. National Library of Medicine, Bethesda, Maryland. www.nlm.nih.gov/research/umls/.

OTA, S., SOGA, M., YAMAMOTO, N. et TAKI, H. (2012). Design and development of a learning support environment for apple peeling using data gloves. *In Proceedings of the Cooking with Computers workshop (CwC)*, pages 7–12.

ROBERT, L. (1990). *Le petit Robert*. Le Robert, Paris.

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

WITTEN, I. et FRANK, E. (2005). *Data mining : Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco.

© ATALA

DEFT2013, une cuisine de caractères

Gaël Lejeune[1] Charlotte Lecluze[1] Romain Brixtel[1]
(1) Équipe HULTECH (GREYC - CNRS UMR 6072 - Université de Caen),
Bd Maréchal juin, 14032 Caen Cedex
prenom.nom@unicaen.fr

RÉSUMÉ

Nous présentons dans cet article les méthodes utilisées par l'équipe HULTECH pour sa participation au DÉfi Fouille de Textes 2013 (DEFT2013). Cette neuvième édition porte sur l'analyse automatique de recettes de cuisine en langue française. Elle comporte quatre tâches : trois de classification de documents et une d'extraction d'information. Notre équipe participe aux quatre tâches. Nous nous appuyons pour chaque tâche sur une technique d'algorithmique du texte : la détection de chaînes de caractères répétées maximales ($rstr_{max}$). Les méthodes développées sont simples et non supervisées.

ABSTRACT

DEFT2013, a distinctive character-based cuisine

We present here the HULTECH (Human Language Technology) team approach for the DEFT2013 (french text mining challenge). The aim of the challenge is to automatically analyze recipes in French. It has four tasks : three of document classification and one of information extraction. Our team participate in four tasks. Our methods relie on a text algorithmics technic : detection of maximal repeated strings ($rstr_{max}$). The developed methods are simple and unsupervised.

MOTS-CLÉS : classification, extraction d'information, appariement, algorithmique du texte.

KEYWORDS: classification, information extraction, matching, text algorithmics.

1 Introduction

Pour cette nouvelle édition du DÉfi Fouille de Textes, quatre tâches d'analyse concernant les recettes de cuisine ont été proposées :

Niveau de difficulté de réalisation d'une recette : La tâche 1 a pour but d'évaluer la capacité d'un algorithme à inférer la difficulté d'une recette de cuisine en se basant sur le texte de la recette et son titre (cf. Figure 1). L'appréciation de la difficulté est donnée sur une échelle à 4 valeurs : très facile, facile, moyennement difficile, très difficile.

Type de plat : À partir du texte et du titre, la tâche 2 propose de classer les recettes en fonction du type de plat préparé, selon trois classes : entrée, plat principal, dessert.

© ATALA

Appariement titre/recette : La tâche 3 consiste à retrouver pour chaque texte de recette, son titre original dans une liste de titres de recettes. Pour chaque texte de recette, le système doit fournir une liste de titres par ordre de pertinence décroissante.

Ingrédients d'une recette : La tâche 4 diffère des précédentes car elle ne concerne pas la classification des recettes, mais l'extraction d'information. Dans cette tâche, il s'agit d'identifier la liste des ingrédients de la recette au moyen d'une liste normalisée globale de libellés d'ingrédients fournie aux participants. Cette liste contient tous les ingrédients présents dans la base des textes de recettes, mais peut aussi contenir des ingrédients qui ne sont pas présents dans les recettes.

2 Description des corpus d'apprentissage et de test

Le corpus utilisé comporte des recettes de cuisine (Figure 1) provenant du site Marmiton. 60% de ce corpus sert pour l'apprentissage, les 40% restant sont utilisés pour le test. La proportion de document de chaque classe entre apprentissage et test est la même. L'équilibre entre corpus d'apprentissage et corpus de test s'arrête là, la taille des fichiers n'ayant pas été vérifiée.

2.1 Corpus d'apprentissage

Le corpus d'apprentissage comprend 13863 documents. Dans le tableau 1, nous présentons la répartition des recettes du corpus dans les différentes classes des tâches 1 et 2.

Type \ Difficulté	Très facile	Facile	Moyennement difficile	Difficile	Total
Entrée	1787 (55,1%)	1259 (38,8%)	189 (5,8%)	10 (0,3%)	3245
Plat principal	3025 (46,9%)	2897 (44,9%)	496 (7,7%)	29 (0,5%)	6449
Dessert	2150 (51,6%)	1595 (38,2%)	383 (9,2%)	41 (1%)	4169
Total	6962	5751	**1068**	**80**	13861 / 13863

TABLE 1 – Répartition des recettes par type de plat et difficulté dans le corpus d'apprentissage. La différence entre les totaux globaux correspond à l'absence de l'étiquette difficulté pour deux plats principaux.

0. http://www.marmition.org

© ATALA

```
<recette id="94727">
<titre>Truites tricolores</titre>
<type>Entrée</type>
<niveau>Moyennement difficile</niveau>
<cout>Moyen</cout>
<ingredients>
<p>2 filets de truite</p>
<p>2 grosses tomates</p>
<p>15g de pignons grillés</p>
<p>1 échalote</p>
<p>zeste de citron</p>
<p>court-bouillon</p>
<p>Pour le pesto:</p>
<p>un gros bouquet de basilic</p>
<p>15-20g de pignons grillês</p>
<p>30-40g de parmesan</p>
<p>1 petite gousse d'ail</p>
<p>20g d'huile d'olive</p>
<p>sel, poivre</p>
</ingredients>
<preparation>
<![CDATA[
Après avoir inciser les tomates, les plonger quelques secondes dans de l'eau frémissante, les ressortir et les peler. Les
couper en deux, les épépiner, les poser sur une plaque huilée (huile d'olive) allant au four, les mettre bien à plat, y poser
dessus l'échalote hachée finement ainsi qu'un peu de zeste de citron, saler, poivrer. Les laisser au four (150°) pour 20
minutes.
En parallèle, plonger les filets de truite dans un court-bouillon tiède [...]
Mixer tous les ingrédients du pesto ou acheter un petit verre de pesto. Mais c'est meilleur fait maison ! Tartiner le pesto
sur les truites. Couvrir le peso avec les lamelles de tomates et manger tout de suite sur des assiettes préchauffées.Décorer
avec les pignons grillés et un filet de court-bouillon ou un filet d'huile d'olive.on peut le prendre comme plat principal en
mettant les portions doubles (2 filets par personne). Un bon verre de Riesling serait un bon accompagnement.
 ]]>
 </preparation>
</recette>
```

FIGURE 1 – Un exemple de recette

2.2 Corpus de test

Un corpus de test différent était fourni pour chaque tâche.

Type \ Difficulté	Très facile	Facile	Moyennement difficile	Difficile	Total
Entrée	298 (52,5%)	234 (41,3%)	34 (6%)	1 (0,2%)	567
Plat principal	477 (45,5%)	471 (44,9%)	91 (8,7%)	9 (0,9%)	1048
Dessert	357 (51,5%)	262 (37,8%)	64 (9,2%)	10 (1,5%)	693
Total	1132	967	189	20	2308

TABLE 2 – Répartition des recettes par type de plat et difficulté, corpus de test de la tâche 1.

© ATALA

Type \ Difficulté	Très facile	Facile	Moyennement difficile	Difficile	Total
Entrée	310	215	34	3	562
Plat principal	518	473	87	5	1083
Dessert	334	251	72	4	661
Total	1162	939	193	12	2306

TABLE 3 – Répartition des recettes par type de plat et difficulté, corpus de test de la tâche 2.

Type \ Difficulté	Très facile	Facile	Moyennement difficile	Difficile	Total
Entrée	283	224	39	0	546
Plat principal	507	495	74	0	1076
Dessert	349	264	65	7	685
Total	1139	983	178	7	2307

TABLE 4 – Répartition des recettes par type de plat et difficulté, corpus de test de la tâche 3.

Type \ Difficulté	Très facile	Facile	Moyennement difficile	Difficile	Total
Entrée	300	166	18	2	486
Plat principal	512	489	77	5	1083
Dessert	392	286	52	6	736
Total	1204	941	147	13	2305

TABLE 5 – Répartition des recettes par type de plat et difficulté, corpus de test de la tâche 4.

3 Description de la méthode et résultats

Pour toutes ces tâches, nous exploitons les mêmes concepts : les chaînes de caractères répétées maximales et les affinités. Nous présentons ci-après les caractéristiques de ces deux concepts.

Ainsi, pour chaque tâche, nous procédons à une recherche de chaînes de caractères répétées maximales, des $rstr_{max}$. Ces chaînes sont déduites d'un tableau de suffixes. Elles sont obtenues en calculant des motifs sans trou tels que décrits par (Ukkonen, 2009) [1]. Ces chaînes possèdent les caractéristiques suivantes :

répétées : les chaînes ont un effectif de 2 ou plus ;

maximales : les chaînes ne peuvent être étendues à gauche ou à droite sans perdre une occurrence.

Le tableau 6 contient les répétitions maximales les plus longues extraite de l'exemple de recette de la figure 1.

1. Les outils permettant le calcul de ces chaînes sont disponibles ici : `https://code.google.com/p/py-rstr-max/`

© ATALA

Nombre de caractères	$rstr_{max}$	Effectif dans la recette
32	'g de pignons grillés '	2
22	' les filets de truite '	2
17	' filets de truite'	3
16	'e court-bouillon'	2
16	' pignons grillés'	3
16	' d'huile d'olive'	2
16	' court-bouillon '	2
15	'zeste de citron'	2
15	'ingredients> '	2
15	' court-bouillon'	3

TABLE 6 – Exemples des $rstr_{max}$ les plus longues dans la recette de la figure 1 « Truite tricolores ».

À l'occasion de notre participation à la campagne DEFT2011 sur l'appariement entre des articles scientifiques et leur résumé , nous avions introduit le concept d'affinité (Lejeune *et al.*, 2011). Une affinité est une $rstr_{max}$ commune à au moins deux documents et qui possède certaines caractéristiques de longueur et d'effectif. La présence d'affinités « hapax » dans la collection d'articles et dans la collection de résumés était caractéristique d'un appariement correct.

Ici, nous utilisons la notion d'affinité pour mettre en relation différentes recettes. Deux recettes partageant un grand nombre d'affinités sont donc considérées comme proches.

À partir de deux recettes du corpus d'apprentissage, nous illustrons la notion d'affinité :

Nombre de caractères	affinité	Effectif
93	'</type> <niveau>Moyennement difficile</niveau>' ' <cout>Moyen</cout> <ingredients> '	2
55	'e </ingredients> <preparation> <![CDATA['	2
34	']]> </preparation> </recette>'	2
25	' gousse d'ail '	2
15	' sel'	2
14	'un peu d'huile'	2
14	' les tomates, '	2

TABLE 7 – Exemples d'affinités entre la recette de la « Truite tricolores » et celle du « Carry de poulet ».

3.1 Tâche 1 : identification du niveau de difficulté

Les recettes de même difficulté ont des contenus proches. Cette similarité entre recette est fondée sur les chaînes de caractères communes entre ces recettes. Nous calculons la similarité entre recettes en utilisant le concept d'affinité décrit précédemment. Ainsi, deux recettes ayant beaucoup d'affinités entre elles possèdent, par définition, beaucoup de $rstr_{max}$. Nous avons

 © ATALA

établi une étiquette par défaut : l'étiquette « Facile ». Au début du traitement, toute recette que nous cherchons à étiqueter reçoit l'étiquette « facile ». La méthode de classification calcule les probabilités d'allocation d'une autre étiquette. Si jamais aucune autre étiquette n'est trouvée, alors la méthode considère la recette comme facile, sinon la nouvelle étiquette est attribuée à la recette.

Partant de cette hypothèse, nous avons établi la chaîne de traitement suivante :

1. Comparaison de la recette à étiqueter avec les recettes dont l'étiquette est connue
2. Classement des appariements par ordre décroissant du nombre d'affinités
3. Sélection des n meilleurs appariements
4. Calculer l'effectif de chaque étiquette dans cette sélection
5. Si une étiquette a un effectif supérieur à $n/2$: elle est choisie [2]

De façon purement arbitraire, nous avons choisi trois valeurs de n : 10, 20 et 50. Pour $n = 10$, le système avait tendance à surgénérer les étiquettes « difficile » et « moyennement difficile ». Pour $n = 50$, il était très rare sur le corpus d'apprentissage qu'une majorité se détache. Sur le corpus de test, cela a même conduit à ce qu'il n'y ait jamais d'autre étiquette attribuée que l'étiquette par défaut « Facile ». Nous avons donc produit une *baseline* naïve qui attribue automatiquement l'étiquette par défaut. Les meilleurs résultats ont été obtenus pour $n = 20$. La plus-value vis-à-vis de la *baseline* restant toutefois très faible.

Run	Mesure	Rappel	Précision	F_1-mesure
# 1	Macro	0.273	0.250	0.261
# 1	Micro	0.489	0.489	0.489
# 2	Macro	0.265	0.248	0.256
# 2	Micro	0.465	0.465	0.465
# 3	Macro	0.289	0.281	0.285
# 3	Micro	0.360	0.360	0.360

TABLE 8 – Résultats sur la tâche d'identification de la difficulté.

3.2 Tâche 2 : identification du type de plat préparé

Pour la tâche d'identification du type de plat, nous avons appliqué une technique très proche de celle utilisée pour la détection de la difficulté. Les classes étant équilibrées, l'étiquette choisie étant cette fois simplement l'étiquette majoritaire sur les n premiers appariements triés par nombre d'affinités.

Nous avons considéré que l'on pouvait discriminer plus facilement les desserts à travers certains termes techniques et ingrédients. À l'opposé, nous avons fait l'hypothèse que les indices pour déterminer qu'une recette était un plat principal était plus rare. En cas d'égalité, nous avons gardé l'étiquette supposée la plus difficile à identifier : entrée *vs* plat ou dessert et plat *vs* dessert.

2. Sinon on conserve l'étiquette par défaut

© ATALA

Run	Mesure	Rappel	Précision	F_1-mesure
#1	Macro	0.760	0.741	0.750
#1	Micro	0.746	0.746	0.746
#2	Macro	0.584	0.609	0.596
#2	Micro	0.573	0.573	0.573
#3	Macro	0.381	0.418	0.398
#3	Micro	0.331	0.331	0.331

TABLE 9 – Résultats sur la tâche d'identification du type de plat.

3.3 Tâche 3 : Appariement du texte d'une recette avec son titre

Pour cette tâche, nous n'avons utilisé le corpus d'apprentissage que pour éprouver le système. Elle est basée sur le nombre d'affinités existant entre une recette et chacun des titres de la base de titres. Nous ne conservons que les activités qui sont hapax dans la base de titres. En cas d'égalité entre plusieurs titres, aucun départage n'est effectué.

Quand un appariement est effectué, on continue le processus en ne tenant plus compte du titre déjà utilisé. Notre idée est qu'à l'issue d'une première phase où les appariements les plus évidents sont effectués, on essaie d'apparier les recettes restantes à un des titres non utilisés.

Il s'est avéré que la première phase donnait de très bons résultats sur le corpus d'apprentissage mais que les résultats se détérioraient rapidement par la suite. Une fois les appariements les plus évidents effectués dans la première phase, les appariements ultérieurs étaient de très mauvais qualité. Notre système de donnait qu'un titre possible par recette, ce qui explique que le rappel et la précision soient rigoureusement identiques (Tableau 10).

Run	Mesure	Rappel	Précision	F_1-mesure
#1	Macro	0.127	0.127	0.127
#1	Micro	0.127	0.127	0.127

TABLE 10 – Résultats sur la tâche d'appariement texte-titre, moyenne réciproque des rangs (*Mean Reciprocal Rank*) : 0.1956.

3.4 Tâche 4 : extraction des ingrédients à partir du titre et du texte de la recette

Pour l'extraction des ingrédients, la méthode est très proche de celle utilisée pour l'appariement titre-recette. On suppose qu'un ingrédient est utilisé dans une recette si l'on trouve une sous-chaîne de cet ingrédient dans le texte. De manière à éviter la surgénération d'ingrédients, un filtrage était opéré : on ne conserve un ingrédient que s'il n'est pas une sous-chaîne d'un autre ingrédient extrait. Par exemple si le système extrait « citron » et « jus de citron », on conserve seulement le second terme. C'est une *baseline* assez naïve, nous l'avons utilisé pour le run1.
Le run2 ajoutait une heuristique que nous avions déduite du corpus d'apprentissage : étant donné f la fréquence d'apparition d'un ingrédient dans la série de recettes, on déduit p la probabilité

 © ATALA

maximale d'apparition du même ingrédient qui est égale à $2.f$. S'il s'avère qu'un ingrédient a une probabilité d'apparition dans le corpus de test supérieure à cet attendu, on introduit des contraintes. L'ingrédient en question n'est sélectionné que s'il apparaît à des positions déterminées : débuts et fins de texte. Le début et la fin du texte de la recette sont les 20 premiers et les 20 derniers caractères. Cette heuristique permettait d'écarter des ingrédients extraits trop fréquemment tels que « dore », « plie », « rose » ou « renne ». Le run 3 remplace cette heuristique par une heuristique sur la longueur des ingrédients, le système n'extrait pas d'ingrédient dont la longueur en caractères est inférieure à 4.

	Run #1	Run #2	Run #3
MAP	0.4881	0.5556	0.5076

TABLE 11 – Moyenne de la précision moyenne (*Mean Average Precision, MAP*) par run sur la tâche d'extraction des ingrédients.

4 Discussion

L'édition 2013 du DÉfi Fouille de Textes proposait l'analyse automatique de recettes de cuisine en langue française. Nous avons présenté pour chaque tâche une méthode fondée sur une analyse au grain caractère. Les résultats que nous avons obtenus ont été décevants puisque nous avons terminé à la dernière place sur chaque tâche à l'exception de la tâche 4 (extraction des ingrédients). Il été intéressant de tester des techniques d'apprentissage en exploitant les caractéristiques issues de notre analyse au grain caractère.

Remerciements

Nous remercions les organisateurs du DEFT2013 d'avoir proposé une tâche originale.

Références

LEJEUNE, G., BRIXTEL, R., GIGUET, E. et LUCAS, N. (2011). Deft2011 : appariement de résumés et d'articles scientifiques fondé sur les chaînes de caractères. *In Défi Fouille de Textes/TALN 2011*, pages 53–64.

UKKONEN, E. (2009). Maximal and minimal representations of gapped and non-gapped motifs of a string. *Theorie in Computer Science*, 410(43):4341–4349.

© ATALA

Systèmes du LIA à DEFT'13

Xavier Bost[1] Ilaria Brunetti[1,6] Luis Adrián Cabrera-Diego[1]
Jean-Valère Cossu[1] Andréa Linhares[1,5] Mohamed Morchid[1]
Juan-Manuel Torres-Moreno[1,2,3,4] Marc El-Bèze[1,2,4] Richard Dufour[1,4]
(1) LIA, 339, chemin des Meinajariès 84912 Avignon Cedex 09
(2) SFR Agorantic Université d'Avignon et des Pays de Vaucluse, 84000 Avignon Cedex
(3) École Polytechnique de Montréal, 2900 Bd Edouard-Montpetit Montréal, QC H3T1J4
(4) Brain & Language Research Institute, 5 avenue Pasteur, 13604 Aix-en-Provence Cedex 1
(5) Universidade Federal do Ceara Rua Estanislau Frota, S/N, CEP 62.010-560 – Sobral/ CE Brésil
(6) Equipe Maestro INRIA, 2004, Route des Lucioles, 06902 Sophia Antipolis
prenom.nom@univ-avignon.fr

RÉSUMÉ

La campagne Défi de Fouille de Textes (DEFT) en 2013 s'est intéressée à deux types de fonctions d'analyse du langage, la classification de documents et l'extraction d'information dans le domaine de spécialité des recettes de cuisine. Nous présentons les systèmes du LIA appliqués à DEFT 2013. Malgré la difficulté des tâches proposées, des résultats intéressants ont été obtenus par nos systèmes.

ABSTRACT

The Systems of LIA at DEFT'13

The 2013 *Défi de Fouille de Textes* (DEFT) campaign is interested in two types of language analysis tasks, the document classification and the information extraction in the specialized domain of cuisine recipes. We present the systems that the LIA has used in DEFT 2013. Our systems show interesting results, even though the complexity of the proposed tasks.

MOTS-CLÉS : Classification de textes ; Extraction d'information ; Méthodes statistiques ; Domaine de spécialité.

KEYWORDS: Text Classification ; Information Extraction ; Statistical Methods ; Speciality domain.

1 Description des tâches et des données

Le Défi Fouille de Textes (DEFT) s'intéresse en 2013 à deux types de fonction d'analyse du langage, la classification de documents (tâches 1 à 3) et l'extraction d'information (tâche 4), ceci dans un domaine de spécialité : les recettes de cuisine. Pour cette nouvelle édition du défi [1], l'équipe du Laboratoire Informatique d'Avignon (LIA) a décidé de participer aux tâches suivantes :

– *T1. L'identification à partir du titre et du texte de la recette de son niveau de difficulté sur une échelle à 4 niveaux : très facile, facile, moyennement difficile, difficile ;*

1. http://deft.limsi.fr/2013/

© ATALA

– *T2. L'identification à partir du titre et du texte de la recette du type de plat préparé : entrée, plat principal, dessert ;*
– *T4. L'extraction à partir du titre et du texte d'une recette de la liste de ses ingrédients.*

Le corpus, fourni par DEFT au format *XML*, est composé de 13 684 recettes. Nous avons choisi d'augmenter notre base de données en extrayant des recettes de cuisine à partir de 4 sites spécialisés : Cuisine AZ (`http://www.cuisineaz.fr`), Madame Le Figaro (`http://madame.lefigaro.fr/recettes`), 750 grammes (`http://www.750g.com/`) et Ptit Chef (`http://www.ptitchef.com/`). Environ 50k recettes supplémentaires ont pu être obtenues à partir de ces sites. Notons cependant que les recettes téléchargées ne comportaient pas toujours les mêmes champs d'information que ceux contenus dans les recettes fournies par les organisateurs de la campagne.

2 Méthodes utilisées

Pour répondre à ce défi, nous avons employé plusieurs méthodes de classification (modèles probabilistes discriminants, *boosting* de caractéristiques textuelles) et d'extraction d'information, ainsi que leurs combinaisons.

2.1 Modèles discriminants

Le système DISCOMP_LIA a été appliqué aux tâches 1 et 2. Son fonctionnement repose essentiellement sur les modèles discriminants décrits dans (Torres-Moreno et al., 2011). Ces modèles servent aussi bien à agglutiner des mots afin de produire des termes composites ayant un fort pouvoir discriminant (sans toutefois passer en dessous d'un seuil préfixé de couverture), qu'à pondérer de façon plus appropriée ces termes dans le calcul de l'indice de similarité (ici Cosine revisitée) entre chaque recette et le sac de mots de chacune des classes (3 ou 4 selon les tâches).

Trois nouveautés ont été introduites pour prendre en compte les particularités du défi :
– nous avons opté pour une approche hiérarchique propre à chacune des 2 tâches :
  – pour la tâche T1 : identification du niveau *facile* ou *difficile* d'une recette, puis dans chacun des 2 groupes, différenciation en 2 sous-groupes (*moyennement* ou *très difficile*) ;
  – pour la tâche T2 : identification du type *dessert* ou *autre* ; dans le cas d'une identification en type *autre*, différenciation entre *plat principal* et *entrée* ;
– nous avons appliqué le même système avec un paramétrage *ad hoc* à 2 jeux de données : le titre de la recette seul pour le premier et pour le second l'ensemble formé par le contenu et le titre de la recette. Une combinaison linéaire entre les scores des 2 systèmes a permis d'obtenir une nouvelle distribution de probabilités sur les hypothèses de classes ;
– la formation par agglutination des termes et ce, surtout dans le titre, a fait apparaître des sous-types de recettes (quiches, soupes, gâteaux) qui ont permis fournir aux modèles de meilleurs points d'appui pour identifier la classe visée. Dans les cas où l'indice de pureté de *Gini* valait 1, et où la couverture était élevée, nous avons enrichi le sac de mots avec des composants factices de façon à compenser le bruit qui pouvait être engendré par le reste de la recette. Ceci a été fait de façon semi-automatique, mais devrait pouvoir être complètement automatisé.

 © ATALA

Notons par ailleurs que nous nous sommes servis du critère de pureté de *Gini* pour sélectionner quelques mots erronés ce qui, après correction, a permis d'augmenter leur couverture sans altérer leur pouvoir discriminant.

2.2 Algorithmes de *boosting*

Pour les tâches de classification (1 et 2), nous proposons une manière de combiner différentes caractéristiques textuelles et numériques au moyen d'un algorithme d'apprentissage automatique : le *Boosting* (Schapire, 2003). Son objectif principal est d'améliorer (de « booster ») la précision de n'importe quel algorithme d'apprentissage permettant d'associer, à une série d'exemples, leur classe correspondante. Le principe général du *Boosting* est assez simple : la combinaison pondérée d'un ensemble de classifieurs binaires (appelés classifieurs *faibles*), chacun associé à une règle différente de classification très simple et peu efficace, permettant au final d'obtenir une classification robuste et très précise (classifieur *fort*). La seule contrainte de ces classifieurs *faibles* est d'obtenir des performances meilleures que le hasard. Dans le cas des classifieurs binaires, l'objectif est de classifier plus de 50 % des données correctement.

Nous proposons d'utiliser l'algorithme *AdaBoost* car il présente de nombreux avantages dans le cadre des tâches de classification proposées. En effet, l'algorithme est très rapide et simple à programmer, applicable à de nombreux domaines, adaptable aux problèmes multi-classes. Il fonctionne sur le principe du *Boosting*, où un classifieur *faible* est obtenu à chaque itération de l'algorithme *AdaBoost*. Chaque exemple d'apprentissage possède un poids, chaque tour de classification permettant de les re-pondérer selon le classifieur *faible* utilisé. Le poids d'un exemple bien catégorisé est diminué, au contraire d'un exemple mal catégorisé, pour lequel on augmente son poids. L'itération suivante se focalisera ainsi sur les exemples les plus « difficiles » (poids les plus élevés). L'algorithme a été implémenté dans l'outil de classification à large-marge *BoosTexter* (Schapire and Singer, 2000), permettant de fournir comme caractéristiques au classifieur des données numériques mais également des données textuelles. Les classifieurs *faibles* sur les données textuelles peuvent prendre en compte des n-grammes de mots. Nous avons utilisé l'outil *IcsiBoost* (Favre et al., 2007), qui est l'implémentation open-source de cet outil. *IcsiBoost* a notamment l'avantage de pouvoir fournir, pour chaque exemple à classifier, un score de confiance pour chacune des classes entre 0 (peu confiant) et 1 (très confiant).

2.2.1 Niveau de difficulté

Pour l'approche utilisant l'algorithme de classification *AdaBoost*, nous avons retenu différentes caractéristiques qui ont été fournies en entrée pour l'apprentissage du classifieur :

- données textuelles
  1. n-grammes de mots du titre de la recette avec taille maximum de 3 mots ;
  2. n-grammes de mots du texte de la recette avec taille maximum de 4 mots ;
  3. uni-gramme sur la liste des ingrédients obtenue automatiquement (voir section 2.6).
- données numériques continues
  1. nombre de mots du titre de la recette ;
  2. nombre de mots du texte de la recette ;

© ATALA

3. nombre de phrases du texte de recette ;
4. nombre de séparateurs du texte de la recette (point, virgule, deux-points, etc.) ;
5. taille de la liste des ingrédients obtenue automatiquement.

À la fin de ce processus d'apprentissage, la liste des classifieurs sélectionnés est obtenue tout comme le poids de chacun d'eux, afin d'utiliser les exemples les plus discriminants pour chaque classe (chaque niveau de difficulté est considéré dans cette tâche comme une classe). Nous avons composé un sous-corpus à partir du corpus d'apprentissage fourni par les organisateurs de la tâche afin d'optimiser le nombre de classifieurs *faibles* nécessaires. Ce sous-corpus est composé de 3 863 recettes respectant la distribution des classes, le reste étant utilisé pour l'apprentissage (environ 72% des données d'apprentissage). Notons que tout le corpus d'apprentissage sera néanmois utilisé pour entraîner les classifieurs pendant la phase de test de la campagne. Enfin, une attention particulière a été portée sur la normalisation des données textuelles du titre et de la description des recettes. En effet, les données textuelles « brutes » fournies par les organisateurs possédaient un vocabulaire très bruité principalement à cause des nombreuses abbréviations (par exemple « th » pour « thermostat ») et fautes d'othographe (par exemple « échalote » et « échalotte »). Quatre traitements particuliers ont été appliqués sur le texte :

1. suppression de la ponctuation et isolation de chaque mot (exemple : « et couper l'oignon. » devient « et␣couper␣l'␣oignon ») ;
2. utilisation d'une liste manuelle de normalisation de mots (exemple : « kg » devient « kilogramme ») ;
3. conversion automatique des chiffres en lettres ;
4. regroupement des *n*-grammes de mots les plus fréquents au sein d'une même entité au moyen de l'outil *lia_tagg* [2] (exemple : « il y a » est considéré comme un mot « il_y_a »).

Ce texte normalisé est utilisé par l'algorithme *AdaBoost* pour la recherche des classifieurs *faibles* sur les données textuelles lors de la phase d'apprentissage, mais également pour la recherche du niveau de difficulté associé à une recette lors de la phase de test. Au final, nous obtenons pour chaque recette à classifier, un score de confiance pour chaque niveau de difficulté, permettant de choisir le niveau de difficulté de la recette (score le plus élevé).

2.2.2 Type de plat

Dans ce problème de classification au moyen de l'approche par *Boosting*, nous avons utilisé les mêmes caractéristiques et normalisations que celles décrites dans la tâche de recherche du niveau de difficulté (voir section 2.2.1). Pour chaque recette, nous obtenons donc un score de confiance pour chaque type de plat qui pourra être utilisé pour choisir la classe à associer à une recette.

2.3 SVMs

Le corpus des recettes $\mathbb{X}$ étant donné, il s'agit d'élaborer à partir d'un sous-ensemble $\mathbb{D} \subset \mathbb{X}$ de recettes annotées par classe (niveau de difficulté ou type de plat selon la tâche), une méthode de

2. http ://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/chercheurs/bechet/download_fred.html

© ATALA

classification γ qui à toute recette associe une classe. En notant $\mathbb{C}$ l'ensemble des classes, nous avons donc :

$$\gamma : \mathbb{X} \longrightarrow \mathbb{C} \tag{1}$$

La troisième des méthodes appliquées a consisté, pour chaque couple de classes $(c, c') \in \mathbb{C} \times \mathbb{C}$ (avec $c \neq c'$), à apprendre et à appliquer un classifieur binaire $\gamma_{(c,c')} : \mathbb{X} \rightarrow \{c, c'\}$ à base de SVMs.

En notant $s_{(c,c')}(r)$ le score obtenu selon le classifieur $\gamma_{(c,c')}$ par la recette $r \in \mathbb{X}$, nous avons alors, avec $c' \in \mathbb{C}$:

$$\gamma(r) = \arg\max_{c \in \mathbb{C}} \left(s_{(c,c')}(r) \right) \tag{2}$$

Lors des phases d'apprentissage et d'application des classifieurs binaires, les recettes ont été représentées vectoriellement dans l'espace du lexique de référence. Le poids $w(t)$ du terme t d'une recette r dans le vecteur associé est donné par :

$$w(t) = tf(t).idf(t) = tf(t).\log\left(\frac{|\mathbb{X}|}{df(t)}\right) \tag{3}$$

où $tf(t)$ désigne le nombre d'occurrences du terme t dans la recette et $df(t)$ le nombre de recettes du corpus $\mathbb{X}$ où le terme t apparaît. L'approche décrite a été appliquée pour les deux tâches de classification.

Dans le cas de la prédiction du type de plat, le lexique a cependant été filtré avant vectorisation des recettes sur le critère de l'information mutuelle entre un terme t et une classe $c \in \mathbb{C}$ (Manning et al., 2008). Après sélection des termes, seuls les 10 000 termes les plus porteurs d'information sur les classes ont alors été retenus.

2.4 Similarité cosinus

La quatrième approche, seulement utilisée pour la prédiction du type de plat, repose elle aussi sur une représentation vectorielle des documents : à chaque recette r, nous faisons correspondre un vecteur v_r dont les composantes $w_r(t)$ dans l'espace du lexique sont données, pour chaque terme t présent dans le corps de la recette, par :

$$\begin{aligned} w_r(t) &= tf(t).idf(t).G(t) \\ &= tf(t).idf(t).\sum_{c \in \mathbb{C}} \mathbb{P}^2(c|t) \\ &= tf(t).idf(t).\sum_{c \in \mathbb{C}} \left(\frac{df_c(t)}{df_{\mathbb{T}}(t)}\right)^2 \end{aligned} \tag{4}$$

où $G(t)$ désigne l'indice de *Gini*, $df_{\mathbb{T}}(t)$ est le nombre de recettes du corpus d'apprentissage $\mathbb{T} \subset \mathbb{X}$ contenant le terme t et $df_c(t)$ correspond au nombre de recettes du corpus d'apprentissage annotées selon le type $c \in \mathbb{C}$.

A chaque classe $c \in \mathbb{C}$, nous faisons par ailleurs correspondre un vecteur dont les composantes $w_c(t)$ sont données dans l'espace du lexique par :

$$w_c(t) = df_c(t).idf(t).G(t) \tag{5}$$

© ATALA

où $df_c(t)$ désigne le nombre de recettes annotées selon le type de plat c où le terme t apparaît.

Une mesure de similarité $s(r,c)$ entre une recette $r \in \mathbb{X}$ et un type de plat $c \in \mathbb{C}$ est alors donnée par le cosinus de l'angle formé par les vecteurs v_r et v_c :

$$s(r,c) = \cos(\widehat{v_r, v_c}) = \frac{\sum_{t \in r \cap c} w_r(t).w_c(t)}{\sqrt{\sum_t w_r(t)^2.w_c(t)^2}} \tag{6}$$

Dans ces conditions la classe $\gamma(c)$ attribuée à une recette r est celle qui maximise la mesure de similarité cosine :

$$\gamma(r) = \arg\max_{c \in \mathbb{C}} (s(r,c)) \tag{7}$$

Le vocabulaire utilisé lors de la phase de vectorisation était constitué des seuls termes t dont l'indice de Gini $G(t)$ était au moins égal à 0.45.

Ce seuil a été fixé empiriquement par maximisation du macro F-score sur un corpus de développement constitué de 3 864 des 13 864 recettes du corpus d'apprentissage.

2.5 Fusion des systèmes

Pour chacune des tâches de classification, plusieurs méthodes ont donc été appliquées : trois pour la tâche 1, quatre pour la tâche 2.

Pour chaque couple $(r,c) \in \mathbb{X} \times \mathbb{C}$ associant une recette du corpus à une classe, nous disposions donc d'un score $s_i(r,c)$ pour chacune des méthodes de classification (avec $i = 1,...,3$ ou $i = 1,...,4$ selon la tâche).

2.5.1 Normalisation des résultats par méthode

Les scores produits en sortie des différents systèmes de classification ont d'abord été normalisés de manière à ce que la somme des prédictions sur l'ensemble des classes pour une méthode donnée soit égale à 1. En notant $n_i(r,c)$ le score normalisé obtenu selon la $i-$ème méthode pour le couple $(r,c) \in \mathbb{X} \times \mathbb{C}$, nous avons donc :

$$n_i(r,c) = \frac{s_i(r,c)}{\sum_{j=1}^{|\mathbb{C}|} s_i(r,c_j)} \tag{8}$$

où c_j désigne la $j-$ème classe ($j = 1,...,4$ ou $j = 1,...3$ selon la tâche de classification).

Deux méthodes de fusion ont été appliquées après normalisation des scores $s_i(r,c)$.

2.5.2 Combinaison linéaire des scores

La première méthode de fusion a consisté, pour une recette donnée r, à départager les classes candidates par combinaison linéaire des scores obtenus selon les différentes méthodes de classifi-

© ATALA

cation. En notant $\gamma(r)$ la classe conjecturée, on a donc :

$$\gamma(r) = \arg\max_{c \in \mathbb{C}} \left(\sum_{i=1}^{m} n_i(r,c) \right) \tag{9}$$

où m correspond au nombre de méthodes appliquées pour la tâche considérée.

2.5.3 Méthode Electre

Issue du domaine de l'optimisation multi-objectif discrète, la méthode Electre (Roy, 1991) consiste à définir sur l'ensemble $\mathbb{C}$ des classes candidates une relation $\mathscr{S} \subset \mathbb{C} \times \mathbb{C}$ dite de surclassement. De manière informelle, nous pouvons dire qu'une classe c surclasse une classe c' si elle la domine sur un nombre « important » de méthodes et si, sur les éventuelles méthodes restantes où elle est dominée par c', elle ne l'est pas au-delà d'un certain seuil fixé pour chaque méthode. Plus formellement :

$$c \mathscr{S} c' \Longleftrightarrow \begin{cases} conc(c,c') \geqslant sc \\ v(c,c') = 0 \end{cases} \tag{10}$$

où $conc(c,c') \in [0,1]$ désigne l'indice de concordance entre les classes candidates c et c', $sc \in [0,1]$ un seuil dit de concordance fixé empiriquement et $v(c,c') \in \{0,1\}$ un indice binaire dit *de veto*.

L'indice de concordance $conc(c,c')$ évalue le taux de méthodes (éventuellement pondérées) selon lesquelles c domine c' :

$$conc(c,c') = \frac{\sum_{i \in M(c,c')} p_i}{\sum_{i \in M} p_i} \tag{11}$$

où M désigne l'ensemble des méthodes, $M(c,c')$ l'ensemble des méthodes sur lesquelles c domine (non strictement) c' et p_i le poids accordé à chaque méthode $i \in M$.

L'indice binaire $v(c,c')$ (à valeurs dans l'ensemble $\{0,1\}$) fixe la valeur du veto de c' vers c selon les modalités suivantes (en notant $n_i(r,c)$ le score obtenu selon la i-ème méthode par le couple (r,c) associant une recette à une classe) :

$$v(c,c') = \begin{cases} 1 \Longleftrightarrow \exists i \in M : n_i(r,c') > n_i(r,c) \text{ et } n_i(r,c') - n_i(r,c) \geqslant v_i \\ 0 \text{ sinon} \end{cases} \tag{12}$$

où $v(i) \in [0,1]$ est la valeur de veto associée à la $i-$ème méthode.

Les valeurs des différents paramètres ont été fixées empiriquement à partir du corpus de développement par maximisation des métriques appliquées pour évaluer la classification.

Les poids p_i associés aux différentes méthodes ont tous été fixés à $p_i = 1$. Le seuil de concordance a été fixé à $sc = 0,7$ pour la tâche 1 et $sc = 0,6$ pour la tâche 2 ; enfin, la valeur de veto v_i a été fixée à $v_i = 0,5$ pour chacune des méthodes et quelle que soit la tâche considérée.

Le noyau de la relation $\mathscr{S} \subset \mathbb{C} \times \mathbb{C}$ est alors constitué des éventuelles classes candidates non surclassées.

© ATALA

Dans le cas d'un noyau *singleton*, la classe conjecturée pour la recette est l'unique classe élément du noyau.

Dans le cas de noyau vide ou de noyau formé de plusieurs classes concurrentes, c'est la meilleure classe candidate selon la méthode de combinaison linéaire qui a été retenue.

2.6 Extraction d'ingrédients

La tâche 4 est une tâche pilote d'extraction d'information. La difficulté est multiple, car les auteurs des recettes n'utilisent pas tous les ingrédients, ou ils ont introduit librement des ingrédients non listés. Par exemple, il n'est pas rare d'avoir des passages comme « mélanger énergiquement tous les ingrédients » (recette 27174), qui ne permettent pas d'extraire directement les ingrédients.

Nous avons attaqué ce problème en utilisant deux approches : l'une à base de règles et l'autre probabiliste. L'idée de base pour les règles a été la suivante : soit A l'ensemble de mots de la recette i ; soit B l'ensemble de mots de la liste DEFT (références d'apprentissage ou *gold-standard* d'évaluation). Éliminer tous les mots de A qui ne sont pas dans B. Cela produit une liste d'ingrédients candidats L_c.

Cette liste peut contenir des termes génériques comme « VIANDE », « FROMAGE » ou « POISSON », présents dans le texte de la recette. Afin de mieux identifier ce terme, nous avons employé une approche probabiliste naïve. Nous avons calculé la probabilité d'avoir un certain type de viande x, étant donné une liste d'ingrédients $L_c = (L_1, L_2, ..., L_n)$:

$$p(x|L_c) = P(x \cap L_c)/p(L_c); \quad x = \{\text{VIANDE, FROMAGE, POISSON}\} \tag{13}$$

Les termes ainsi identifiés, sont alors injectés dans la liste extraite L_c, ce qui produit la liste deifnitive L.

La liste d'ingrédients L ainsi obtenue a été utilisée dans les systèmes de *boosting* (c.f. section2.2) dans les tâches T1 et T2.

3 Résultats et discussion

Les systèmes ont été évalués en utilisant les mesures proposées par DEFT : le F-score pour les tâches 1 et 2, et la mesure *Mean Average Precision* (MAP) de TREC pour la tâche d'extraction.

3.1 Tâches de classification

Pour la tâche de classification en niveau de difficulté (T1), nos trois runs étaient donnés par :

- Run 1 : SVMs seuls ;
- Run 2 : fusion des trois méthodes par ELECTRE ;
- Run 3 : combinaison linéaire des trois méthodes.

Pour la tâche de classification en type de plat (T2) :

- Run 1 : Modèle discriminant seul ;

© ATALA

– Run 2 : fusion des quatre méthodes par ELECTRE ;
– Run 3 : combinaison linéaire des quatre méthodes.

Les résultats obtenus sont récapitulés dans les tableaux 1 (tâche 1) et 2 (tâche 2).

Pour la tâche 1 et pour chacun des trois runs, nous fournissons, outre les F-scores, la distance moyenne (micro-écart) de l'hypothèse à la référence sur l'échelle des quatre niveaux de difficulté, deux niveaux contigüs étant distants de 1.

T1	**run 1**	**run 2**	**run 3**
Micro F-score	**0.5916**	0.5812	0.5877
Macro F-score	0.4531	0.4531	**0.4569**
Distance moyenne	**0.4409**	0.4586	0.4465

TABLE 1 – *Résultats obtenus sur le corpus de test - Tâche 1*

T2	**run 1**	**run 2**	**run 3**
Micro F-score	0.8760	0.8860	**0.8886**
Macro F-score	0.8706	0.8795	**0.8824**

TABLE 2 – *Résultats obtenus sur le corpus de test - Tâche 2*

Pour la tâche 2, la combinaison des méthodes permet d'obtenir des scores systématiquement supérieurs à ceux que nous obtenons en appliquant une méthode isolée.

Pour la tâche d'évaluation du niveau de difficulté, les bénéfices retirés de l'utilisation d'une méthode de fusion sont moins nets et apparaissent dépendants de la métrique appliquée lors de l'évaluation.

Ceci a permis au LIA de se positionner dans la tâche 1 : 3ème/6 et dans la tâche 2 : 1er/5.

En ce qui concerne la tâche T2, nous croyons que l'évaluation aurait dû prendre en compte la dimension muti-labels de plusieurs recettes. Par exemple la recette 18 052 (présente dans le test) se termine par « servir tiède (aussi bon chaud que froid ou réchauffé), en entrée ou en plat » est annotée uniquement comme une entrée. Si un système produit l'étiquette *plat principal*, faut-il pour autant compter cela comme une erreur ?

Relevons aussi l'aspect subjectif de la tâche T1. En effet, dire qu'une recette est facile ou difficile sans tenir compte qui en est l'auteur c'est faire fi de son niveau d'expertise ! ! !

Nous émettons également l'idée que les macro-mesures devraient être privilégiées aux micro-mesures car ces dernières pousseraient à mettre en œuvre des stratégies négligeant les classes minoritaires. Nous pensons donc que cette façon d'évaluer serait plus adaptée à une utilisation orientée application.

© ATALA

3.2 Tâche d'extraction

L'évaluation de cette tâche est assez subjective, malgré son caractère d'extraction d'information. Par exemple, dans l'ensemble d'apprentissage, la recette 10 514 dit : « 40 cl d'Apremont ou autre blanc de Savoie sec (facultatif, mais donne plus de goût) ». Les références DEFT indiquent **mais** comme ingrédient de la tartiflette. Evidemment nous n'avons pas incorporé l'information des ingrédients d'apprentissage, à la différence de la méthode DEFT ayant généré le *gold-standard*.

Nous avons obtenu un score de MAP=0.6364 en mesure *qrel* dans le corpus d'apprentissage, et dans le classement officiel un score de MAP=0.6287 dans le corpus de test. Cela a positionné l'équipe du LIA au rang 3ème/5. Bien que standard, la mesure MAP pourrait avoir intégré les accents dans l'évaluation des ingrédients extraits : cela aurait permis de lever des ambiguïtés qui ont été inutilement introduites par la désaccentuation.

Nous pensons que les références fournies par DEFT dans l'ensemble d'apprentissage, ont contribué à rendre la tâche floue. En effet, cela n'a pas de sens d'extraire *papier sulfurisé* ou *couteau* (extraite à partir de *pointe de couteau*) comme ingrédients. De même, l'ingrédient *maïs*, incorrectement extrait des recettes d'apprentissage comportant l'expression *mais...*, n'est pas apparu dans les recettes de test comportant ladite expression.

3.3 Comparaison versus les humains

Nous avons mis en place une petite expérience impliquant des êtres humains. Pour la tâche T4, nous avons demandé à 7 annotateurs (experts ou non en matière culinaire), d'effectuer la tâche DEFT. Ceci nous a permis de positionner nos systèmes par rapport aux performances des personnes. Puisque les personnes ont des connaissances extra-linguistiques, les juges humains ont eu comme seule consigne celle de ne pas consulter des ressources externes (par exemple des ressources électroniques ou des livres de cuisine) pour classer les recettes ou pour extraire les ingrédients. Le tableau 3 montre la moyenne MAP pour chaque annotateur A_i sur 50 recettes de la tâche T4, choisies au hasard. La moyenne générale pour les personnes est de MAP=0.5433 et pour le système S de MAP=**0.6570**. Ceci montre que dans ce sous-ensemble, nos systèmes sont au-dessus des performances atteintes par les humains.

A1	A2	A3	A4	A5	A6	A7	S
0.5333	0.6029	0.5271	0.5880	0.5313	0.4679	0.5529	**0.6570**

TABLE 3 – Performance des personnes sur la tâche T4.

Il est clair que, même les humains ayant une expertise culinaire, ont obtenu de piètres notes dans cette tâche. Cela s'explique par la mauvaise qualité du *gold standard* fourni par DEFT. Par exemple, la crème ou le café (cuillère à café, recette 90 806) semblent subir une extraction de nature aléatoire... D'ailleurs, nous avons souvent été surpris par les ingrédients de référence de DEFT utilisés dans telle ou telle recette, qui ne peuvent en aucun cas avoir été employés de façon conjointe. Par exemple, dans la recette 64 761 (dessert), la référence DEFT liste parmi les ingrédients « moules » et « caramel » ! Évidemment il s'agit de l'ustensile *moule à charlotte*. Malheureusement il s'agit de problèmes récurrents. Sans parler de l'eau ou de la pâte (déclinée comme pâté, pâtes fraîches, brisée, sablée, etc.) qui semblent être détectées aléatoirement. Pour répondre au défi, nous avons développé plusieurs systèmes d'extraction d'ingrédients qui ont dû

 © ATALA

être modifiés (souvent à l'encontre de ce qui nous semblait logique) afin de rendre une sortie proche de celle des références. Ceci revient en fin de compte à modéliser la machine d'extraction de DEFT. Les références étant de qualité discutable, est-il intéressant de chercher à reproduire la sortie d'une telle machine ?

4 Conclusions

Malgré la quantité d'erreurs présentes dans le corpus d'apprentissage, nos algorithmes ont conduit à des résultats intéressants. Nous voulons ajouter quelques commentaires par rapport à ce défi. Nous avons observé que l'auto-évaluation du niveau de difficulté des recettes par ceux qui les publient est un exercice très subjectif, assez souvent sujet à caution. En ce qui concerne le type de plat, l'évolution du mode de vie rend assez floue la distinction entre plat principal et hors d'œuvre. Les tartes salées, les quiches ou les tartines sont de plus en plus considérées comme des plats « de résistance ». Ceci conduit souvent les internautes à exprimer de façon explicite leur indécision quant à la catégorie à retenir. Ces observations n'ont pas suffi à atténuer notre perplexité lorsque nous avons constaté que certaines méthodes à base de « cuisine » algorithmique réussissaient à faire mieux que des approches sophistiquées. Pour autant, nous n'avons pas cédé à la tentation de les retenir pour en faire une soumission !

Remerciements

Nous remercions les annotateurs qui ont bien voulu nous aider dans cette tâche ! Patricia Velázquez, Sulan Wong, Mariana Tello, Alejandro Molina et Grégoire Moreau. Nous tenons aussi à remercier les organisateurs de la campagne d'évaluation, sans qui nous n'aurions pu participer à ce défi.

Références

Favre, B., Hakkani-Tür, D., and Cuendet, S. (2007). Icsiboost. `http://code.google.com/p/icsiboost`.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Roy, B. (1991). The outranking approach and the foundations of ELECTRE methods. *Theory and Decision*, 31 :49–73.

Schapire, R. E. (2003). The Boosting Approach to Machine Learning : An Overview. In *Nonlinear Estimation and Classification*.

Schapire, R. E. and Singer, Y. (2000). BoosTexter : A boosting-based system for text categorization. In *Machine Learning*, volume 39, pages 135–168.

Torres-Moreno, J., El-Bèze, M., Bellot, P., and F., B. (2011). *Peut-on voir la détection d'opinions comme un problème de classification thématique ?*, chapter 9. Hermes Lavoisier.

© ATALA

Approches hybrides pour l'analyse de recettes de cuisine DEFT, TALN-RECITAL 2013

Luca Dini, André Bittar, Mathieu Ruhlmann
Celi France, 12-14 rue Claude Génin, 38000 Grenoble
dini@celi-france.com, bittar@celi-france.com, ruhlmann@celi-france.com

RÉSUMÉ

Le Défi fouille de textes (DEFT) 2013 porte sur l'analyse automatique de recettes de cuisine en langue française, une thématique actuellement à la mode et qui a déjà fait l'objet d'une campagne d'évaluation (Computer Cooking Contest, 2012). Dans cet article, nous présentons nos modules pour le traitement de la tâche 1 (la détection du niveau de difficulté d'une recette) et de la tâche 4 (l'extraction des ingrédients d'une recette). Notre système pour la tâche 1 repose sur une combinaison de méthodes d'apprentissage et de logique probabiliste, alors que pour la tâche 4, nous avons utilisé des méthodes hybrides (symboliques et statistiques). Nous posons certaines questions concernant le traitement automatique d'un corpus qui contient des données subjectives, ainsi que la pertinence de la mesure MAP pour l'évaluation de la tâche 4.

ABSTRACT

Hybrid approaches for the analysis of cooking recipes

The Défi fouille de textes (DEFT) 2013 focuses on the automatic processing of cooking recipes in French, a topic that has already been the subject of an evaluation campaign (Computer Cooking Contest, 2012). In this article, we present our systems for the processing of tasks 1 (identifying the level of difficulty of a recipe) and 4 (extracting ingredients from a recipe). Our system for task 1 is based on a combination of machine learning and probabilistic logic, while for task 4, we used hybrid (symbolic and statistical) methods. We raise certain questions concerning the automatic processing of a corpus containing subjective data, as well as the appropriateness of using the MAP measure for the evaluation of task 4.

MOTS-CLÉS : anaylse de recettes, extraction d'information, classification automatique, méthodes hybrides
KEYWORDS : recipe analysis, information extraction, automatic classification, hybrid methods

1 Introduction

Le Défi Fouille de Textes (DEFT) 2013 porte sur l'analyse automatique de recettes de cuisine en langue française. Cette thématique, actuellement à la mode, a déjà fait l'objet d'une campagne d'évaluation, sous forme du « Computer Cooking Contest » (CCC) (Computer Cooking Contest, 2012) tenu tous les ans depuis 2008. Le but de l'édition 2012 de la campagne CCC était d'encourager l'utilisation de moyens technologiques pour aider à la création de recettes de cuisine. Il s'agissait d'une tâche de *génération* de recettes.

En revanche, cette édition du DEFT consiste en quatre tâches d'*analyse* de recettes de

© ATALA

cuisine :

1. Identifier à partir du titre et du texte de la recette son niveau de difficulté sur une échelle à 4 niveaux : très facile, facile, moyennement difficile, difficile.
2. Identifier à partir du titre et du texte de la recette le type de plat préparé : entrée, plat principal, dessert.
3. Apparier le texte d'une recette à son titre.
4. Extraire du titre et du texte d'une recette la liste de ses ingrédients.

Nous présentons, dans cet article, notre système d'analyse linguistique HOLMES, et nos travaux pour le traitement des tâches 1 et 4.

2 HOLMES

HOLMES (Hybrid Operable platform for Language Management and Extensible Semantics) est une plateforme pour l'analyse de textes écrits en français, italien et anglais, développée au sein de Celi France. C'est avec certains modules de ce système que nous avons abordé les tâches 1 et 4 du DEFT 2013.

HOLMES permet de faire les traitements TAL de base (détection de phrases, tokenisation, étiquetage morphosyntaxique) ainsi que des traitements de plus haut niveau, tels que l'analyse syntaxique en dépendances, l'identification d'entités nommées et l'extraction d'information. La plateforme permet également l'entraînement de classifieurs pour l'apprentissage. HOLMES est développé entièrement en Java et a été conçu pour maximiser son extensibilité pour des vraies tâches de TAL.

Techniquement, HOLMES est une plateforme pour le traitement automatique des langues basée sur une approche *radicalement incrémentale*. Cela veut dire que toutes les informations ajoutées par le système restent accessibles à travers les différentes étapes de traitement. Actuellement, les principaux traitements disponibles dans HOLMES sont les suivants :

- Segmentation en phrases
- Tokenisation
- Etiquetage morphosyntaxique
- Analyse morphologique
- Règles de patrons linéaires sur des séquences d'objets linguistiques
- Analyse syntaxique en dépendances
- Interrogation de base ontologique
- Extraction d'entités nommées (avec Conditional Random Fields)
- Classification automatique

Le module de règles sur des séquences d'objets linguistiques, que nous avons utilisé pour la tâche 4, s'inspire du module TokensRegexp (Chang, 2011) du package CoreNLP développé à l'Université de Stanford (Stanford, 2010). Ces règles permettent l'application d'annotations sur des séquences d'objets linguistiques (ex. tokens). Dans ces règles, tous les traits linguistiques d'un objet donné peuvent être utilisés, ce qui résulte en un système de règles assez puissant. La figure 1 donne un exemple d'une telle règle. Cette règle s'applique sur une séquence commençant par un token ayant le lemme « filet », suivi du token « de » ou « d' », suivi d'un token étiqueté comme nom commun ou nom propre. La règle applique

© ATALA

l'annotation « INGREDIENT » sur la séquence rencontrée dans un texte.

```
{ pattern : ( [{lemma :filet}] /de|d'/ [{tag:/NC|NPP/}] ),
  action : ( Annotate($0,"ner","INGREDIENT") ) }
```

FIGURE 1– exemple de règle sur séquence d'objets linguistiques.

La nature hybride de HOLMES est apparente dans l'intégration étroite de méthodes symboliques (des règles linguistiques écrites à la main par des experts) avec des méthodes basées sur apprentissage. En effet, la sortie des modules statistiques est accessible au niveau des modules symboliques, et inversement, les modules statistiques peuvent recevoir comme traits d'entrée les résultats des calculs symboliques.

3 Tâche 1 : détecter le niveau de difficulté d'une recette

3.1 Description de la tâche 1

La tâche 1 du DEFT 2013 consiste à déterminer, pour une recette de cuisine donnée en entrée, son niveau de difficulté sur une échelle à 4 niveaux : très facile, facile, moyennement difficile, difficile. Pour cette tâche, un système peut se baser sur toutes les informations qu'il est possible d'extraire à partir du texte de la recette et son titre. La mesure d'évaluation pour cette tâche est la distance entre la réponse du système et la réponse correcte.

Malgré sa simplicité apparente, cette tâche s'avère très intéressante dans la mesure où elle implique un certain type de classification, notamment une classification *fonctionnelle,* tout comme la classification en genre (Karlgren & Cutting, 1994) (Kessler et al., 1997). Dans une tâche de classification thématique « standard », les mots du texte caractérisent son appartenance à une classe ou une autre, ce qui se prête à une approche de classification par « sac de mots ». En revanche, pour une tâche de classification fonctionnelle, ce genre de corrélation n'est pas toujours utile car les critères de bonne classification se trouvent dans les *relations* entre mots. Par exemple, deux recettes qui décrivent le même dessert avec deux niveaux de difficulté différents sont thématiquement plus proches qu'une recette de plat principal et une recette de dessert ayant le même degré de difficulté.

Pour cette tâche, un corpus d'apprentissage a été fourni aux participants. Ce corpus contient un total de 13 865 recettes provenant du site web Marmiton.org et qui sont associées à leur niveau de difficulté. Pour chaque recette, l'appréciation de sa difficulté est donnée par son auteur.

3.2 Approche : classification à deux niveaux

L'approche que nous avons adoptée pour aborder cette tâche est basée sur deux niveaux de classification. Le premier niveau repose sur l'entraînement d'un classifieur basé purement sur le corpus d'apprentissage. Le deuxième niveau consiste en l'application de formules logiques qui confirment ou contredisent les résultats de la première étape de classification, permettant de les réviser. Les deux classifieurs utilisent le même ensemble de traits. Etant donnée la nature fonctionnelle de cette tâche, la sélection des traits, que nous présentons dans la section 3.2.1, a été primordiale.

© ATALA

3.2.1 Sélection de traits

Pour cette tâche, la sélection de traits doit être guidée par la question « qu'est-ce qui rend une recette difficile ? ». Intuitivement, on pourrait considérer la longueur du texte comme un facteur décisif, un postulat basé sur la supposition implicite que plus une recette est longue, plus le nombre d'instructions nécessaire à son accomplissement est important. Si cette supposition peut s'avérer vraie pour un livre de recettes « normalisées », ce n'est pas le cas pour un recueil de recettes générées par une communauté d'internautes. En effet, un auteur peut ajouter à une recette courte, une longue explication de ses origines, des souvenirs personnels concernant sa découverte, des faits intéressants sur les ingrédients, etc., comme c'est le cas du corpus de cette tâche. Pour un corpus contenant de telles recettes, la simple longueur du texte n'est pas un critère approprié et nous l'avons donc exclu come trait dans notre traitement.

Une autre approche serait d'adopter comme trait le nombre d'occurrences de verbes trouvés dans le texte. Typiquement, un verbe dénote une action, et donc une opération à effectuer dans la recette. Plus il y a de verbes, plus il y a d'actions à effectuer, et plus la recette est compliquée et difficile. Cependant, nos expérimentations avec cette approche n'ont donné que des résultats moyens.

Une approche plus raisonnée serait alors d'admettre que, dans le corps de recettes publiées par une communauté d'internautes, peuvent apparaître des verbes qui dénotent des actions liées au domaine de la cuisine, mais aussi des verbes qui sont en dehors de ce champ sémantique et qui ne doivent pas être pris en compte. Dans le domaine des « prédicats de cuisine » on peut faire la distinction entre des prédicats qui dénotent des actions intrinsèquement faciles, telles que « faire chauffer » et « faire bouillir », et des prédicats qui dénotent des actions qui demandent plus d'expérience, telles que « flamber ». Bien que cette distinction ne soit pas pertinente pour un classifieur sac-de-mots (qui apprend de façon autonome les prédicats qui caractérisent les recettes faciles et difficiles), elle peut être très importante pour un classifieur manuellement paramétré. Nous avons donc opté pour cette approche.

Afin de différencier les prédicats « faciles » des prédicats « difficiles » et « neutres » nous avons compté le nombre d'occurrences de chacun des verbes dans les recettes des différents niveaux de difficulté. Un prédicat appartient à la classe pour laquelle il a le plus grand nombre d'occurrences. Nous aurions pu adopter des méthodes plus fines, telles que « log odds ratio » (Everitt, 1992), pour établir ce genre de correspondance, mais pour des raisons de simplicité de mise-en-œuvre, nous avons opté pour l'approche que nous venons de décrire. Avec cette méthode, nous avons produit une liste de prédicats associés à un seul niveau de difficulté.

La seule détection des prédicats de cuisine est susceptible, dans certains cas, de produire des erreurs dans les cas où le prédicat est modifié par un syntagme prépositionnel ou adverbial. Par exemple, les instructions « mélangez le tout », « mélangez longuement le tout » et « mélangez longuement le tout avec attention » dénotent des actions avec, dans l'ordre, un degré de complexité croissant. Dans notre calcul de traits pour apprentissage, nous supposons que plus un verbe a de modifieurs, plus l'action qu'il dénote est complexe. Bien

 © ATALA

entendu, un modifieur peut n'impliquer aucune augmentation dans la complexité d'une action (ex. « déposez le mélange **dans un bol** »), mais une vérification manuelle a montré que cela n'est pas très fréquent dans le corpus utilisé pour cette tâche. Nous avons repéré les modifieurs de verbes à l'aide d'une analyse syntaxique en dépendances et en extrayant la tête syntaxique de syntagmes gouvernés par un prédicat donné.

Une recette peut parfois contenir des assertions sur sa difficulté. Il s'agit, dans ce cas, d'une information qui relève du domaine de l'analyse d'opinion plutôt que de la classification pure et, en tant que telle, nécessiterait un traitement indépendant. Cependant, nous n'avons pas disposé de suffisamment de temps pour mettre en place un système pour analyser les assertions subjectives sur la difficulté des recettes. Nous avons alors simplement étiqueté tous les mots selon leur appartenance à un domaine de difficulté ou son opposé. Nous avons constitué un lexique de ces mots qui dénotent une polarité facile ou difficile. Au moment de l'exécution, les adjectifs de cette liste sont convertis en leur équivalent adverbial.

Pour résumer, les traits que nous avons utilisés pour entraîner notre classifieur sont les suivants :

- Classes de verbes selon le niveau de difficulté d'une action liée à la cuisine
- Nombre de modifieurs des verbes
- Classes de mots indiquant une polarité facile ou difficile

3.2.2 Classifieur simple

La première étape de classification consiste en l'entraînement d'un classifieur d'entropie maximale (MaxEnt) sur l'ensemble de traits mentionné dans la section 3.2.1. Afin d'éviter que le classifieur ne s'oriente vers la tâche de classification thématique (à la place d'une classification fonctionnelle), nous n'avons pas fourni en entrée le texte entier. Pour la même raison, nous n'avons pas non plus fourni la liste des ingrédients détectés, qui était pourtant rendue disponible par la tâche 4. La liste de prédicats « faciles » et « difficiles » repérés dans le texte, ainsi que la liste de mots renvoyant à une polarité facile ou difficile détectés, ont été fournies en entrée au classifieur. Dans cette configuration, nous avons obtenu une « précision »[1] d'environ 67%.

3.2.3 Réseaux logiques de Markov

Le deuxième niveau de classification que nous avons effectué consiste en l'application de réseaux logiques de Markov (Richardson, 2006). Un réseau logique de Markov (RLM) est une logique probabiliste qui permet la modélisation de problèmes qui peuvent être décrits en termes de prédicats et formules logiques, et qui permet de faire des inférences incertaines (Beedkar et al., 2013). Les RLM ont été utilisés pour obtenir des résultats satisfaisants dans d'autres tâches de TAL, tels que la désambiguïsation sémantique et l'étiquetage de rôles sémantiques (Che & Ting, 2010) (Riedel & Meza-Ruiz, 2008) ou l'analyse sémantique non supervisée (Poon & Domingos, 2009). Nous avons utilisé l'implémentation des RLM fourni dans le package open-source Tuffy (Niu et al., 2011). Une des motivations pour avoir choisi d'utiliser les RLM réside dans le fait que les packages tels que Tuffy permettent d'apprendre

[1] Nous utilisons la formule simple suivante pour calculer la « précision » : *éléments correctement détectés / éléments corrects*. Il ne s'agit donc pas de la mesure de précision habituellement utilisée en extraction d'information.

© ATALA

automatiquement, à partir d'un corpus de référence, les poids à attribuer aux prédicats. Par exemple, prenons les deux formules suivantes :

1. Recipe(x) & Contains(x, butter) & Has_difficult_preds(x, y) && y > 10 → DIFF(x)
2. Recipe(x) & Contains(x, butter) & Has_difficult_preds(x, y) && y < 10 → EASY(x)

Le système calcule les probabilités de 1 et 2 sur un corpus de référence contenant les prédicats cibles DIFF(x) et EASY(x). L'intérêt de la pondération de prédicats vient du fait qu'un système, suivant l'approche de (Riedel, 2008), peut générer un ensemble très important de prédicats possibles et décider lesquels mèneraient à une bonne conclusion.

Malheureusement, étant donné le corpus de référence à notre disposition, notre méthode de détection automatique du niveau de difficulté a donné des résultats médiocres. En effet, si nous avons obtenu quelques résultats prometteurs (40% de précision pour une échelle de difficulté à 4 niveaux, et 60% pour une échelle à 2 niveaux de difficulté), une inspection manuelle des poids attribués aux prédicats a montré que l'algorithme a résulté en un surapprentissage, sans montrer de réelle connexion avec l'appréciation humaine de la difficulté des recettes.

Après cette phase d'expérimentation, nous avons alors décidé d'omettre la pondération automatique sur la base du corpus de référence pour favoriser des règles écrites à la main. Nous avons écrit des règles pour distinguer les recettes sur une échelle de difficulté bipartite – facile et difficile – la distinction sur une échelle à 4 niveaux étant en dehors de la portée de notre système. Nous donnons ci-dessous quelques exemples des types de règles que nous avons écrites. Pour une recette *r* donnée :

1. Si *r* contient plus de *x* prédicats « difficiles » et plus de *y* modifieurs → DIFF(*r*)
2. Si *r* contient plus de *x* prédicats et contient le mot « long » ou contient le mot « difficile » → DIFF(*r*)
3. Si le ratio de mots à polarité « facile » aux mots de polarité « difficile » > *y* → EASY(*r*)
4. Si le ratio de mots à polarité « difficile » aux mots de polarité « facile » > *y* → DIFF(x)
5. Si le nombre d'ingrédients identifiés > x et le nombre de prédicats > y → DIFF(x)
6. Si l'ingrédient *i* est détecté et que *r* contient au moins un mot à polarité « difficile » → DIFF(x)

L'application de ces règles a encore entraîné une légère baisse de précision. Mais encore, un examen manuel des résultats a montré que ces erreurs étaient souvent dues à des imprécisions dans le corpus de référence où une fausse appréciation avait été attribuée. Sans prendre en compte ces erreurs, les règles manuelles donnaient une nette amélioration des résultats.

3.2.4 Approche hybride

L'approche finale que nous avons adoptée pour cette tâche se compose de deux parties :

1. Un classifieur d'entropie maximale (MaxEnt) entraîné sur des prédicats pertinents et des mots renvoyant aux 4 niveaux de difficulté
2. Un classifieur RLM (à base de règles) capable de distinguer uniquement entre des recettes faciles et difficiles

© ATALA

Comme la tâche demande une classification sur 4 niveaux de difficulté, nous avons gardé le classifieur MaxEnt comme décideur principal et nous sommes servis du classifieur RLM pour effectuer un ré-ranking des résultats. Dans le cas où le classifieur RLM attribuait la catégorie « facile », les scores du classifieur MaxEnt pour les catégories « facile » et « très facile » se voyaient augmenter. Inversement, si le classifieur RLM attribuait la classe « difficile », les scores MaxEnt pour « difficile » et « très difficile » étaient revus à la hausse.

L'amélioration apportée par l'utilisation du classifieur RLM était minime, de l'ordre de 1% de précision (une augmentation de 67% à 68%). Cependant, après une inspection manuelle des résultats, nous avons décidé de garder ce module dans le système car les erreurs étaient dues, pour la plupart, à une « mauvaise » évaluations de la part de l'auteur de la recette dans le corpus de référence.

3.3 Discussion: mes recettes ne sont jamais difficiles!

Sur l'ensemble du corpus, nous avons constaté que l'ensemble des recettes « difficiles » et « moyennement difficiles » ne constitue qu'environ un dixième (1 148) des recettes des classes « facile » et « très facile » (12 714). Cette différence émerge-t-elle simplement du fait d'une prévalence « objective » de recettes faciles, ou relève-t-elle plutôt d'un effet des intentions des auteurs ? Nous tentons, par la suite, de répondre à cette question.

Nous estimons que le concept de la difficulté d'une recette de cuisine peut se définir de deux façons. Dans le cadre d'une publication contrôlée, il existe des critères définis par l'équipe éditoriale pour déterminer la difficulté. Dans ce cas-là, une certaine homogénéité est imposée aux auteurs. Cependant, dans le cas de recettes éditées par des utilisateurs, l'évaluation de la difficulté se base sur trois critères :

1. Les capacités culinaires de l'auteur
2. L'intention de l'auteur : souhaite-t-il que sa recette devienne populaire auprès du grand public ou plutôt qu'elle reste au sein d'une communauté de « passionnés de la cuisine » ?
3. L'interface graphique

Contrairement à d'autres tâches d'évaluation, par exemple l'analyse d'opinions (cf. DEFT 2007 et 2009), les appréciations de la difficulté des recettes sur le site Marmiton.org ne représentent pas une sorte d' « intelligence collective » de gens qui ont essayé et évalué les recettes. Dans notre cas, le seul juge est l'auteur de la recette lui-même. Nous estimons que les 3 points que nous mentionnons ci-dessus, que nous allons examiner de plus près par la suite, représentent un fort biais pour le niveau « facile ».

Concernant les capacités culinaires, un sentiment de fierté peut conduire les auteurs à favoriser la note « facile ». Lorsqu'un auteur évalue sa recette comme étant « difficile », il ou elle s'expose implicitement à être considéré comme un cuisinier moyen par d'autres qui considéreraient la recette « facile ». Par conséquent, les auteurs auraient tendance à attribuer l'appréciation « facile » pour éviter de ternir leur réputation.

Il existe également certains indices qui révèlent l'intention de l'auteur quant au public visé. Un parcours un peu spéculatif du site révèle, par exemple, la recette de « quenelles de volaille, sauce financière », généralement considérée comme difficile. L'auteur a évalué cette recette comme étant facile, malgré la suite élaborée d'instructions qu'il a décrites et qui sont

 © ATALA

manifestement non triviales à accomplir. Les commentaires sur cette recette, laissés par des internautes, témoignent du fort désaccord qui peut exister entre les différents jugements :

> **louve_2:** *C'est bon, simple et rapide*
>
> **laure04130:** *Bien compliqué pour un résultat insatisfaisant.*

La subjectivité du jugement se retrouve également dans le fait que l'on trouve deux versions d'un même plat marquées avec des niveaux de difficulté différents. Par exemple, on trouve « bouillabaisse facile et sa rouille » étiqueté « facile » et « bouillabaisse comme à Marseille » annoté « difficile ». En lisant la recette, on se rend compte que cette différence d'appréciation ne s'explique pas par une différence dans les séquences d'instructions (qui sont, en fait, comparables pour ces deux recettes). On peut supposer qu'il s'agit plutôt d'une position « idéologique » que chacun des auteurs adopte dans la présentation de sa recette : dans le premier cas « accessible à tous » et, dans le deuxième, « recette traditionnelle régionale ».

Enfin, lors de la saisie de l'appréciation par l'auteur, la valeur par défaut dans l'interface est « facile ». Il ne semble pas irraisonnable de supposer que parfois un auteur, que ce soit par paresse ou inadvertance, ne change pas cette valeur par défaut.

En vue de ces observations, nous concluons que ces trois critères rendent difficile l'évaluation de la difficulté de recettes de cuisine sur la base d'éléments purement textuels, au moins sur un corpus comme celui de Marmiton.org.

3.4 Conclusions sur la tâche 1

Nous avons présenté une approche pour la détection du niveau de difficulté de recettes de cuisine basée sur un mélange de méthodes d'apprentissage et de logique probabiliste. Les résultats préliminaires, obtenus sur une partie de 20% du corpus non utilisée pour l'apprentissage, sont assez décevants. De tels résultats médiocres pourraient s'expliquer par un mauvais choix d'algorithmes de classification. Cependant, une analyse manuelle du corpus d'évaluation nous laisse penser que l'évaluation du niveau de difficulté d'une recette est une tâche hautement subjective et que le résultat ne peut être déterminé de façon fiable sur la base du simple texte de la recette. Il serait intéressant de vérifier cette impression par une analyse approfondie des données structurées du site Marmiton.org qui prendrait en compte non seulement les recettes, mais aussi les utilisateurs (les auteurs et les gens qui laissent des commentaires). Nous proposons, si une expérience analogue à la tâche 4 de DEFT 2013 était proposée, que la classification de la recette prenne en compte un paramètre supplémentaire, notamment l'utilisateur et ses comportements.

4 Tâche 4 : détecter les ingrédients d'une recette

4.1 Description de la tâche 4

La tâche 4 du DEFT 2013 concerne l'identification des ingrédients dans les recettes de cuisine. Il s'agit d'une tâche d'extraction d'information dans laquelle un système doit fournir en sortie une liste des ingrédients repérés dans le titre et le texte d'une recette donnée en entrée. L'évaluation s'effectue en comparant la liste d'ingrédients obtenue par le système avec celle fournie par l'auteur de la recette. La mesure de « Mean Average Precision » (MAP) est utilisée pour l'évaluation.

© ATALA

4.2 Corpus et ressources fournis

Le corpus utilisé pour cette tâche (et pour les autres tâches de cette campagne) provient du site web Marmiton (Marmiton). Ce site, ouvert gratuitement au grand public, recense plus de 55 000 recettes de cuisine proposées par ses membres. Les recettes ont été extraites automatiquement du site pour constituer un corpus de test et un corpus d'apprentissage. Les organisateurs ont veillé à ce que les proportions des classes de recettes présentes dans les deux corpus soient les mêmes, mais sans vérifier que les fichiers de chaque corpus soient de la même taille. Pour la tâche 4, le corpus d'apprentissage contient 13 864 fichiers. Le corpus de test, quant à lui, contient 2 306 fichiers.

Une liste globale normalisée contenant (au moins) tous les ingrédients présents dans le corpus global (et éventuellement d'autres qui n'y figurent pas) a également été fournie. La figure 2 montre un extrait de cette liste.

```
2965 0 thon 1
3401 0 aneth 1
3401 0 carre-frais 1
3401 0 citron 1
3401 0 pate-feuilletee 1
3401 0 poivre 1
```

FIGURE 2 – extrait de la liste d'ingrédients.

4.3 Une approche hybride

L'approche que nous avons adoptée pour cette tâche repose sur une combinaison de méthodes symboliques (à base de règles) pour l'extraction d'information et de techniques statistiques souvent utilisées dans les domaines de la recherche d'information et de l'apprentissage. Notre choix de méthodes a été motivé par les facteurs suivants :

1. Il s'agit d'une tâche typique du domaine de l'extraction d'information et, étant donnée la liste d'ingrédients fournie par les organisateurs, dans un monde idéal, le recours à des méthodes uniquement symboliques permettrait d'obtenir des résultats très satisfaisants ;
2. Cependant, le contenu du corpus de test n'a pas été filtré et il reste certains cas où la détection correcte des ingrédients est impossible sur la simple base du texte lui-même. Dans ces cas, et en vue de l'évaluation par la mesure MAP, il est toujours mieux de fournir une réponse « raisonnable » que de ne proposer aucun ingrédient. Les méthodes statistiques ont l'avantage de permettre ce type de robustesse.

Les cas que nous mentionnons au point 2 apparaissent dans les situations suivantes :

- Lorsque la préparation consiste simplement à mélanger les ingrédients, sans qu'ils ne soient mentionnés dans le texte lui-même. Par exemple, « mixer le tout », « mélanger tous les ingrédients », « mélanger l'ensemble », etc. Ce sont des cas de coréférence nominale pour lesquels on ne peut trouver l'antécédent sans considérer la liste d'ingrédients et le corps de la recette comme un texte uni.
- Lorsqu'un ingrédient figure dans la liste des ingrédients avec son nom complet, mais

© ATALA

qu'il est repris dans le corps de la recette par un terme plus générique. Par exemple, la liste d'ingrédients contient « filet de rascasse » alors que le corps de la recette mentionne « les filets » ou « le poisson », encore des cas de coréférence nominale qui sont souvent difficiles à résoudre de façon automatique[2].

- Lorsqu'un ingrédient est implicite dans une des opérations nécessaire au déroulement de la recette. L'auteur présuppose que le cuisinier saura à quel moment utiliser cet ingrédient. Par exemple, la liste d'ingrédients contient « huile d'olive » qui est inclus dans la recette dans l'instruction « faites revenir les oignons ».

Dans la section 4.3.1, nous décrivons les méthodes symboliques utilisées pour traiter les mentions « explicites » d'ingrédients. Dans la section 4.3.2, nous présentons le traitement des cas plus « incertains » par des méthodes statistiques. Finalement, nous présentons, à la section 4.4, nos conclusions sur cette tâche.

4.3.1 Traitement symbolique

Pour la tâche 4 du DEFT 2013, nous avons employé deux couches de traitement symboliques : une couche lexicale et une couche de règles. Pour la couche de traitement lexical, nous avons d'abord construit un lexique de toutes les expressions contenues dans la liste d'ingrédients fournie par les organisateurs. Les entrées du lexique sont représentées sous forme d'expressions régulières Java. Pour chaque élément de la liste d'origine, notre lexique stocke la version lemmatisée de cet élément.

Les entrées lexicales permettent un match « flou », c'est-à-dire que la recherche ne prend en compte ni la casse des caractères, ni les accents (tous les accents dans une position donnée sont considérés comme corrects). La figure 3 illustre la composition de notre lexique.

```
thon                        INGREDIENT
aneth                       INGREDIENT
carr[eéèëê] frais           INGREDIENT
citron                      INGREDIENT
p[aâäà]te feuillet[eéèëê]e  INGREDIENT
poivre                      INGREDIENT
```

FIGURE 3 – extrait du lexique des ingrédients.

La recherche lexicale constitue la première étape du traitement. Les expressions trouvées dans le texte sont alors étiquetées `INGREDIENT`.

Après la recherche lexicale, s'applique la couche de règles sur des patrons linéaires de tokens. Nous avons implémenté des règles pour les situations suivantes :

- **Normalisation de la forme lemmatisée :** le système repère les ingrédients dans le texte sur la base du lemme. Cet ensemble de règles détecte et normalise les cas où les ingrédients doivent apparaître au pluriel dans la sortie, par exemple « marrons », « olives », « lasagnes ».
- **Expansion de termes génériques :** repérage d'ingrédients absents de la liste « gold

[2] En effet, la résolution des relations de coréférence constitue une tâche TAL à part entière (MUC-6, 1995) (MUC-7, 1997).

© ATALA

standard » sous forme simple, mais qui peuvent être utilisés pour expansion de termes. Par exemple, le mot « sauce » n'est pas un ingrédient possible en soi. Cependant, il figure parfois dans le corps des recettes, par exemple, dans une référence générique à un ingrédient de la liste d'ingrédients. Ces règles font l'expansion du mot « sauce » pour retrouver la forme complète la plus plausible dans le contexte.

- **Expansion de formes contractées :** pour trouver une forme complète. Par exemple, « chantilly » donne « creme-chantilly ».
- **Extraction d'ingrédients « implicites » :** pour dériver les noms d'ingrédients à partir de certains verbes. Par exemple, « fariner » et « farine », « poivrer » et « poivre », « saler » et « sel », etc.
- **Correction de l'étape de recherche lexicale :** pour corriger l'application trop simpliste des lexiques. Par exemple, dans l'expression « noix de beurre », seul « beurre » (et pas « noix ») est un ingrédient.

A la fin des étapes de traitement symbolique, la sortie de notre système consiste en une liste d'ingrédients « sûrs ».

4.3.2 Traitement statistique

Le fait d'utiliser MAP comme mesure d'évaluation laisse la place à l'incertitude dans les résultats (du fait de demander une liste ordonnée de possibilités). Paradoxalement, avec cette mesure, un système qui ne liste que des ingrédients 100% sûrs (des ingrédients correctement détectés dans un texte) aurait de moins bons résultats qu'un système qui énumérerait également tous les autres ingrédients possibles. Une partie importante de l'expérience est de dresser une liste ordonnée des ingrédients « incertains » selon leur plausibilité.
Pour enrichir la liste de résultats fournie par les règles, nous avons employé trois heuristiques.
La première a consisté à faire l'expansion de termes génériques. Par exemple, certaines expressions d'ingrédients génériques sont employées à la place de formes plus spécifiques. Tableau 1 ci-dessous donne quelques exemples.

Forme générique	Formes spécifiques
sauce	sauce-tomate, sauce-soja, sauce-worcestershire
porc	roti-de-porc, echine-de-porc, cotes-de-porc

TABLE 1 – Exemples d'expressions d'ingrédients génériques et de formes spécifiques correspondantes.

Une liste de formes complètes est générée à partir de la liste d'ingrédients fournie par les organisateurs en faisant l'expansion des termes qui apparaissent soit au début d'une mention d'ingrédient (ex. « farine » dans « farine-de-ble »), soit à la fin (ex. « agneau » dans « cote-d-agneau »). Les expansions obtenues sont ensuite ordonnées selon leur fréquence

© ATALA

dans le corpus de référence. Il est important de noter que nous n'avons disposé d'aucune ontologie ou taxonomie sur le domaine en question. Par conséquent, certains types d'expansion n'ont pas pu être opérés. Par exemple, l'expansion de « poisson » en une liste de tous les types de poissons (« cabillaud », « morue », « perche », etc.), ou bien l'expansion de « légume » en « chou », « carotte », « potiron », etc.

La deuxième heuristique, correspondant également à une expansion de termes, vise à identifier les ingrédients dans les recettes avec des instructions peu informatives, telles que « mélangez tous les ingrédients ». Pour effectuer cette étape, nous avons utilisé une autre information sur les recettes, notamment le type de plat préparé (entrée, plat principal ou dessert). Un simple classifieur d'entropie maximale nous a permis d'obtenir des résultats de 88% de précision pour la tâche de classification de recettes selon le type de plat préparé. Il est clair que la liste d'ingrédients possibles est en partie restreinte par le type de plat en question. Il est très peu probable qu'un dessert contienne une épaule d'agneau ou qu'un plat principal contienne de la crème Chantilly. Sur la base de cette distinction, nous avons réordonné tous les ingrédients en fonction de leur fréquence d'apparition dans les différents types de plats. Lors du traitement (après toutes les autres étapes d'extraction et d'enrichissement), nous avons classifié la recette en fonction du type de plat et enrichi la liste d'ingrédients possibles avec les ingrédients les plus probables pour ce type de plat.

La troisième et dernière heuristique était d'ajouter tous les autres ingrédients selon leur ordre d'apparition dans le corpus.

4.4 Conclusions sur la tâche 4

Le système que nous avons développé pour cette tâche repose sur un lexique fourni par les organisateurs, des règles écrites manuellement et spécifiques au domaine, et un ensemble d'heuristiques pour trouver des ingrédients « implicites ». Il est clair que l'approche est biaisée pour accommoder la mesure d'évaluation MAP, qui privilégie le rappel sur la précision. Alors que l'utilisation de cette mesure permet la comparaison des résultats avec d'autres campagnes qui utilisent la même mesure, nous estimons que la MAP n'est pas la mesure optimale pour une tâche d'extraction d'information pure. En outre, il est clair que l'utilisation de la mesure MAP, adaptée dans un contexte académique, ne donne pas d'idée réelle de l'utilité de l'application. En effet, nous ne connaissons aucune application réelle pour laquelle il est plus avantageux de fournir un ensemble de faits « plausibles » que de fournir une liste de faits « sûrs ».

Références

Marmiton. (s.d.). Consulté le 5 6, 2013, sur MARMITON: http://www.marmiton.org

Computer Cooking Contest. (2012). Consulté le 5 6, 2013, sur Computer Cooking Contest: http://computercookingcontest.net

BEEDKAR, K., DEL CORRO, L., & GEMULLA, R. (2013). Fully Parallel Inference in Markov Logic Networks. *BTW*, 205-224.

CHANG, A. X. (2011, 9 14). Consulté le 5 6, 2013, sur Tokens Regexp: http://nlp.stanford.edu/software/tokensregex.shtml

© ATALA

CHE, W., & TING, L. (2010). Jointly Modeling WSD and SRL with Markov Logic. (A. f. Linguistics, Éd.) *Proceedings of the 23rd International Conference on Computational Linguistics*, 161-169.

EVERITT, B. (1992). The analysis of contingency tables (Vol. 45). Champman & Hall/CRC.

KARLGREN, J., & CUTTING, D. (1994). Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. *Proceedings of the 15th Conference on Computational Linguistics*. 2, pp. 1071-1075. Association for Computational Linguistics.

KESSLER, B., NUMBERG, G., & SCHÜTZE, H. (1997). Automatic Detection of Text Genre. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 32-38). Association for Computational Linguistics.

MUC-6. (1995). Coreference Task Definition. *Proceedings of the Sixth Message Undertanding Conference*, (pp. 335-344).

MUC-7. (1997). Coreference Task Definition. *Proceedings of the Seventh Message Understanding Conference*, 3.

NIU, F., RE, C., DOAN, A., & SHAVLIK, J. (2011). Tuffy: scaling up statistical inference in Markov logic networks using an RDBMS. *Proceedings of the VLDB Endowment*, 4(6), 373-384.

POON, H., & DOMINGOS, P. (2009). Unsupervised Semantic Parsing. (A. f. Linguistics, Éd.) *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 1, 1-10.

RICHARDSON, M. D. (2006). Markov Logic Networks. *Machine Learning*, 62(1-2), 107-136.

RIEDEL, S., & MEZA-RUIZ, I. (2008). Collective Semantic Role Labelling with Markov Logic. (A. f. Linguistics, Éd.) *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, 193-197.

RIEDEL, S. (2008). Improving the Accuracy and Efficiency of Map Inference for Markov Logic. *UAI '08: Proceedings of the Annual Conference*.

STANFORD. (2010). Stanford CoreNLP: A Suite of Core NLP Tools. Consulté le 5 6, 2013, sur http://nlp.stanford.edu/software/corenlp.shtml

© ATALA

Participation d'Orange Labs à DEFT2013

Olivier Collin, Aleksandra Guerraz, Yannick Hiou, Nicolas Voisine
Orange Labs
2, avenue Pierre Marzin, 22307 Lannion Cedex, France
{olivier.collin,aleksandra.guerraz}@orange.com
{yannick.hiou,nicolas.voisine}@orange.com

RÉSUMÉ

Cet article décrit notre participation à DEFT2013. Nous détaillons tout d'abord notre contribution aux tâches 1 et 2 qui sont des problèmes de classification de recettes de cuisine : classification en niveaux de difficulté (*facile, difficile…)* pour la tâche 1 et vers une typologie (*entrée, plat principal, dessert*) pour la tâche 2. Nous poursuivons ensuite par le traitement de la tâche 3 dont le but était de mettre en correspondance les titres des recettes en regard de leur description. Pour cette tâche, nous avons testé l'utilisation d'une distance sémantique entre mots, distance pré-calculée à partir de Wikipédia français.

ABSTRACT

Orange Labs participation to DEFT 2013

This paper describes our participation to DEFT2013. On the one hand we first detail our contribution to the tasks 1 and 2 which are classification tasks applied to recipes. Task 1 focused on the difficulty level of each recipe *(facile, difficile…)* and task 2 on the natural typology of recipes: *entrée, plat principal, dessert*. On the other hand, the goal of task 3 was to find the link between a recipe description and a title list. For this task, we applied a semantic distance between words, pre-calculated on Wikipedia French corpus.

MOTS-CLÉS : classification supervisée, co-clustering, distance sémantique, Wikipédia, Khiops, TiLT
KEYWORDS : supervised classification, co-clustering, semantic distance, Wikipedia, Khiops, TiLT

1 Introduction

Le Défi Fouille de Texte 2013 proposait 4 tâches, nous avons participé aux tâches 1, 2 et 3. Les deux premières tâches ont été traitées par des algorithmes de classification supervisée. La tâche 3 a été traitée par des calculs de similarité textuelle s'appuyant sur une métrique de distance entre mots. La tâche 4 n'a pas été traitée. Si les tâches 1 et 2 sont plutôt classiques, la nature des données et la confusion des étiquettes ont apporté une difficulté supplémentaire à ce type de tâche. Nous pensons avoir apporté un éclairage intéressant sur la tâche 2 par l'utilisation d'un outil de co-clustering permettant une factorisation utile des variables. En ce qui concerne la tâche 3, nous avons choisi d'explorer une métrique basée sur une distance sémantique inter mots introduite par Cilibrasi (Cilibrasi et Vitanyi 2006, 2007), (Cilibrasi *et al.*, 2009). Nous nous sommes placés dans un cadre applicatif que nous avons repris dans le challenge « AI Mashup Challenge » (Belaunde *et al.*, 2013). Ce cadre consiste à rapprocher un document (recette) d'un ensemble de titres issus de données vidéos (présentation vidéo d'une recette).

 © ATALA

Dans la section 2, nous décrivons le corpus et les prétraitements généraux que nous avons effectués pour les trois tâches. Notre contribution aux tâches 1 et 2 est présentée dans la section 3. La section 4, quant à elle, détaille l'approche que nous avons utilisée pour la tâche 3. Nous terminons enfin par une discussion sur les résultats et une conclusion sur notre participation au défi.

2 Prétraitements généraux

Le corpus d'entrainement est constitué de 13864 recettes de cuisine. Les recettes sont au format XML. Pour chaque recette, les champs « id », « titre », « type », « niveau », « coût », « ingrédients » et « préparation » sont renseignés. De plus, un fichier récapitulatif de ce corpus a été fourni. Il regroupe des informations collectées dans chaque recette ; on y trouve pour chaque recette un champ comportant les ingrédients normalisés (sans accent, en minuscule, en multi-mots) et un champ composé du nombre de ces ingrédients normalisés. Au final, nous avons conservé chacun des champs des recettes mais nous n'avons pas utilisé les ingrédients normalisés. Nous n'avons pas non plus utilisé le fichier fournissant la liste des titres séparément des recettes.

Outre le prétraitement linguistique que nous avons effectué pour la tâche 3 (et décrit en tâche 3), nous avons dû, tout d'abord, traiter l'encodage des recettes. Cet encodage annoncé comme étant de l'UTF8 était ponctuellement bruité notamment par un encodage parasite de caractères HTML. L'encodage des données XML issues du WEB et notamment des flux RSS est très souvent bruité. Nous avons d'autre part converti en minuscules l'ensemble des données des corpus d'apprentissage et de test.

3 Tâche 1 et 2, généralités

Pour ces tâches, nous avons testé deux méthodes de classification différentes : une méthode utilisant le principe de maximum d'entropie et une autre qui exploite l'hypothèse Bayesienne naïve. Rappelons que la tâche 1 consistait à identifier à partir du titre et du texte de la recette son niveau de difficulté sur une échelle à 4 niveaux : *très facile, facile, moyennement difficile, difficile*. Quant à la tâche 2, elle consistait à classifier la recette selon le type de plat parmi *entrée*, *plat principal* et *dessert*. Des premiers essais de classification nous ont permis d'analyser les caractéristiques de chacune de ces tâches. La tâche 1 se caractérise par un étiquetage subjectif de la notion de difficulté d'un plat. Un calcul d'accord inter-annotateur aurait probablement montré une forte divergence des étiquettes. D'autre part, le corpus d'apprentissage est fortement déséquilibré et contient relativement peu d'exemples de la classe *difficile*. La tâche 2 se caractérise aussi par une confusion importante portant sur les classes *entrée* et *plat principal* alors que la classe *dessert* est plutôt bien différenciée des deux autres classes. Dans les deux cas, on se retrouve donc assez loin d'un corpus d'apprentissage idéal.

Concernant ces deux tâches, le premier essai nous a permis de réaliser une référence en utilisant une technique standard de type « Maximum d'Entropie » (Nigam *et al.*, 1999), notre but étant ensuite de tester un outil développé par Orange Labs : Khiops (Boullé, 2008). Pour réaliser notre classifieur de type Maximum d'Entropie, nous avons simplement téléchargé et adapté les entrées/sorties du code mis à disposition par (Tsuruoka *et al.*, 2009) à l'adresse http://www.nactem.ac.uk/tsuruoka/maxent/.

© ATALA

Pour ces deux tâches, pour les autres essais, nous avons donc utilisé Khiops, un outil de préparation des données et de modélisation pour l'apprentissage supervisé et non-supervisé. Cet outil est basé sur une méthode de classification qui exploite l'hypothèse Bayesienne naïve (Boullé, 2007). Cette méthode estime les probabilités conditionnelles univariées. Elle effectue des discrétisations et groupements de valeurs optimaux pour les variables numériques et catégorielles. Elle recherche un sous-ensemble de variables consistant avec l'hypothèse Bayesienne naïve, en utilisant un critère d'évaluation selon une approche Bayesienne de la sélection de modèles et des heuristiques efficaces d'ajout/suppression de variables. Enfin, elle moyenne l'ensemble des modèles évalués en utilisant un lissage logarithmique de la distribution a posteriori des modèles. L'outil Khiops ne nécessite aucun paramétrage, ce qui constitue son point fort. Cet outil, développé par Orange Labs, est utilisé intensivement pour de nombreux problèmes d'analyse et de classification de données mais moins souvent sur des données textuelles. Il nous a permis de réaliser la classification supervisée des documents pour les tâches 1 et 2, ainsi qu'une analyse des variables pour chacun des modèles générés. Il offre, d'autre part, un algorithme de co-clustering que nous avons appliqué à des regroupements documents/mots. Ces regroupements sont à la fois utiles par la sémantique portée mais constituent aussi de nouvelles variables synthétiques permettant de réaliser, voire améliorer les tâches de classification supervisée. Khiops fournit, d'autre part, un ensemble d'outils d'analyse des variables qui permettent notamment d'ordonner et visualiser le poids relatif de chacune d'entre elles.

Nous n'avons utilisé aucun prétraitement de type linguistique (lemmatisation…) pour ces deux tâches. Par contre, différents types de variables ont été utilisés conjointement : des variables catégorielles comme le « coût » ainsi que des variables numériques comme le nombre de mots du champ « préparation ».

Pour pré-évaluer ces tâches, nous avons partagé le corpus d'apprentissage fourni en deux sous-corpus : 70% pour l'apprentissage et 30% pour le test.

Pour la tâche 1, nous observons un fort déséquilibre du nombre d'exemples par classe, la classe *difficile* étant particulièrement sous représentée.

3.1 Tâche 1

3.1.1 Essai 1

L'essai 1 utilise le classifieur de type « Maximum d'Entropie ». Les variables utilisées sont les variables catégorielles « coût » (*bon marché, moyen, assez cher*) et « type » (*entrée, plat principal, dessert*). Nous prenons également en compte comme variables les mots qui sont présents dans le titre, la liste d'ingrédients et dans la préparation. Certaines catégories de mots ont été filtrées en utilisant notamment la liste de « stop words » fournie par NLTK (http://nltk.org/). Ce sont notamment les articles, prépositions, pronoms personnels, adjectifs possessifs, négations, conjonctions, auxiliaires être et avoir. Nous avons en plus supprimé la plupart des termes représentant des valeurs numériques, des symboles de mesure de contenance, de poids… Les petits mots ne contenant qu'une ou deux lettres ont été aussi éliminés. Les variables numériques utilisées sont le « nombre de mot dans la préparation » et le « nombre d'ingrédients ». Ces variables ont été discrétisées linéairement.

© ATALA

Les résultats sur notre sous-corpus de test sont très mauvais pour cette tâche : toutes les recettes sont classées comme *facile*, soit un taux d'erreur de 66%.

Les fréquences relatives des classes sont probablement la cause de ce problème. Nous ne nous attendions donc pas à de bons résultats pour cet essai.

3.1.2 Essai 2

Pour la tâche 1 utilisant Khiops (essai 2), nous utilisons les variables catégorielles suivantes : « coût » (*bon marché, moyen, assez cher*) et « type » (*entrée, plat principal, dessert*). Les variables numériques utilisées sont : « nombre de mot dans la préparation » et « nombre d'ingrédients ». Nous prenons également en compte comme variables les mots qui sont présents dans le titre de la recette et dans sa préparation. Les variables « mots » liées aux ingrédients n'ont pas apporté une information utile pour cette tâche avec cet outil. Ce choix de variables a été réalisé suite à l'analyse de variables fournie par Khiops.

Les graphes suivants produits par l'environnement de Khiops illustrent cette analyse pour les variables les plus pertinentes pour notre tâche.

Nous observons ainsi une corrélation entre le niveau de difficulté d'une recette et son coût (figure 1). Cette figure illustre le fait que lorsqu'une recette est « *assez cher* » il est moins probable qu'elle soit très facile à réaliser.

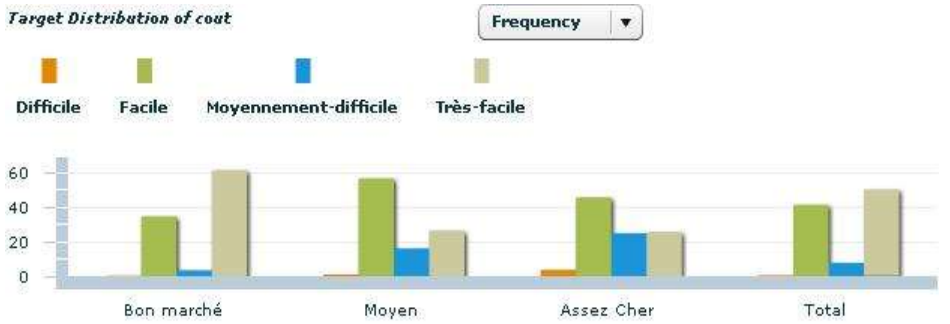


FIGURE 1 – Distribution des valeurs de la variable « coût ».

Le nombre de mots de la préparation est aussi une variable importante (figure 2).

© ATALA

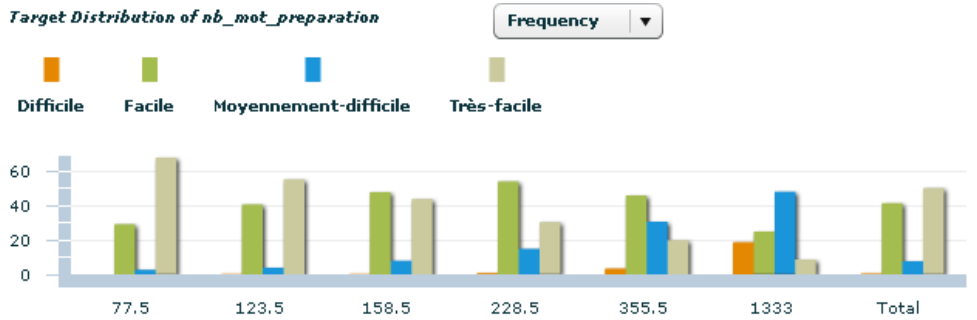


FIGURE 2 – Distribution des valeurs de la variable « nombre de mots de la préparation ».

Plus la recette est longue, plus la difficulté de la réalisation augmente.

Parmi les varibles sélectionnées par Khiops, nous trouvons des mots comme « poche » et

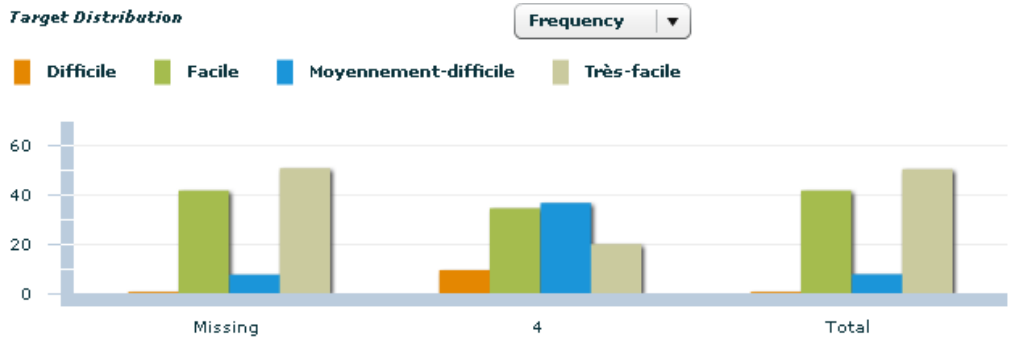


FIGURE 3 – Distribution pour le mot « poche ».

« salade ». Les figures 3 et 4 présentent la distribution de ces mots. Le mot « poche » est absent de la plupart de recettes. En revanche, quand il apparaît il est significatif : il est plus probable que la recette soit difficile à réaliser. Dans le corpus le mot « poche » apparaît dans les contextes « poche à douille » ou « poche à encre », il s'agit, en effet, de termes techniques. En ce qui concerne le mot « salade » (figure 4), il est plus probable qu'une recette soit facile à réaliser quand le « mot » salade apparaît.

© ATALA

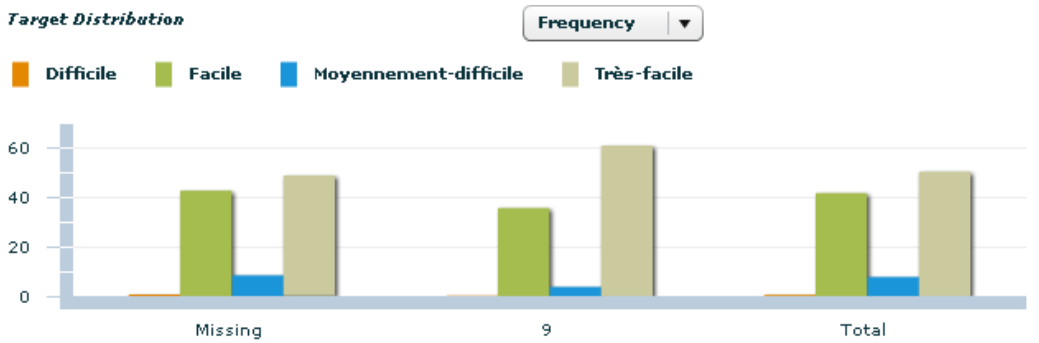


FIGURE 4 – Distribution pour le mot « salade ».

Outre la performance atteinte, cet environnement d'analyse s'avère très utile dans la compréhension du problème. Les tests de mise au point ont montré que Khiops proposait un résultat intéressant pour cette tâche. Mais, même si cet outil souffre moins du manque de données associées à la classe *difficile* et propose une modélisation et une analyse de variables pertinentes, la tâche reste délicate.

3.2 Tâche 2

3.2.1 Essai 1

L'essai 1 utilise le classifieur de type « Maximum d'Entropie », prenant à peu près les mêmes variables d'entrée que pour la tâche 1. Bien évidemment, la variable « type » n'est pas présente, mais nous avons utilisé la variable « niveau». Pour la mise au point de cet essai, nous avons constitué un corpus d'apprentissage équi-réparti contenant 2268 exemples par classe, soit environ 50% des exemples disponibles. Les tests préliminaires ont donc été réalisés sur les exemples restants.

Contrairement à la tâche 1, les résultats sont utilisables. Nous avons notamment observé que des mots comme « tarte » et « quiche » sont « confondus » pour les classes *plat principal* et *entrée*. Nous avons constaté que la classe *dessert* est plutôt bien discriminée, un *dessert* étant confondu avec un *plat* ou une *entrée* moins de 1% des cas. La confusion provient essentiellement des classes *entrée* et *plat principal*, une *entrée* étant confondue avec un *plat principal* presque une fois sur deux, un *plat principal* étant parfois confondu avec une *entrée* dans près de 8% des cas.

Pour notre test final, nous avons donc conservé une équi-répartition des données d'apprentissage de 3200 exemples par classes.

3.2.2 Essai 2

Pour la tâche 2 utilisant Khiops (essai 2), nous utilisons les variables catégorielles « coût » et « niveau » ainsi que les variables numériques « nombre de mot dans la préparation » et « nombre d'ingrédients », ainsi que les variables « mots » présents dans le titre et dans la préparation.

Pour cette tâche, nous utilisons Khiops pour une modélisation non-supervisée et supervisée.

© ATALA

Nous avons donc aussi recours à sa fonctionnalité de co-clustering (Boullé, 2011). Les deux dimensions que nous utilisons sont les textes, et les mots. Après convergence de l'algorithme, nous obtenons 102 clusters de textes et 219 clusters de mots. Nous utilisons ensuite les clusters de textes pour la modélisation supervisée. Les recettes ne sont plus représentées par un ensemble de mots, mais sont classées dans un des 102 clusters. Nous utilisons également les variables catégorielles « coût » et « niveau » pour la modélisation supervisée et les variables numériques. Cependant ces variables n'ont pas de gros impact sur le modèle. La variable « cluster » a un pouvoir prédictif beaucoup plus significatif que les autres variables. L'analyse des regroupements de documents réalisés par le co-clustering permet d'expliquer ce résultat : les regroupements effectués de manière non-supervisée correspondent globalement à des regroupements d'entrées, de plat et de desserts. Si nous examinons plus finement les clusters de textes, nous observons en réalité une spécialisation par « cluster » dont voici quelques exemples :

Les recettes *Gâteau pomme-rhubarbe crousti-moelleux, Gâteau automnal, Gâteau aux fruits jaunes et aux pignons, Clafoutis corrézien* appartiennent au même « cluster », caractérisant un dessert de type gâteau. Nous trouvons également des clusters de desserts à base de fruits ou encore à base de chocolat.

De la même manière, les entrées *Velouté de printemps, Soupe aux marrons et aux carottes, Crème de céleri glacée à la menthe, Crème de cresson au caviar de Sevruga, Soupe d'endive à la bière* ont été regroupées au sein d'un même cluster qui semble correspondre à des entrées chaudes de type soupe ou crème.

Un autre exemple concerne des entrées à base de fruits de mer ou de poissons : *Pamplemousse de la mer, Mini-concombres farcis du pêcheur, Miettes de crabe et pamplemousse, Brochette des îles, Salade d'été au saumon fumé, feta, pommes et carottes…*

Pour les plats principaux, nous avons par exemple un cluster contenant : *Coq au vin de la mère Michèle, Bolognaise façon Mag, Bœuf mijoté façon bourguignon, Garbure landaise, Daube de canard au madiran, Daube de sanglier à la provençale…*

Finalement, examinons un cluster qui contient des entrées et des plats principaux et montre la difficulté de dissocier les classes *entrée* et *plat principal* : *Quiche à la truite fumée, aux courgettes et à la mozzarella (plat), Tarte à l'aubergine et à la brousse (entrée), Quiche à la tomate et aux oignons (plat), Quiche chorizo et tomate (entrée), Tarte moutardée au thon, tomates et poireaux (entrée), Tarte légère courgettes, jambon et chèvre gratiné (entrée), Tarte lardons champignons camembert (plat), Tarte à la tomate et au roquefort (plat), Quiche au poulet et à l'estragon (plat), Quiche au parmesan et tomates (entrée), Quiche au crabe (entrée).* On retrouve dans ces titres et dans ces recettes les termes « quiche » et « tarte » qui conviennent à un plat ou une entrée.

La figure 5 illustre la distribution des classes cibles (*entrée, plat principal, dessert*) pour un regroupement de clusters réalisé par Khiops. On visualise assez nettement la spécialisation réalisée par le co-clustering.

© ATALA

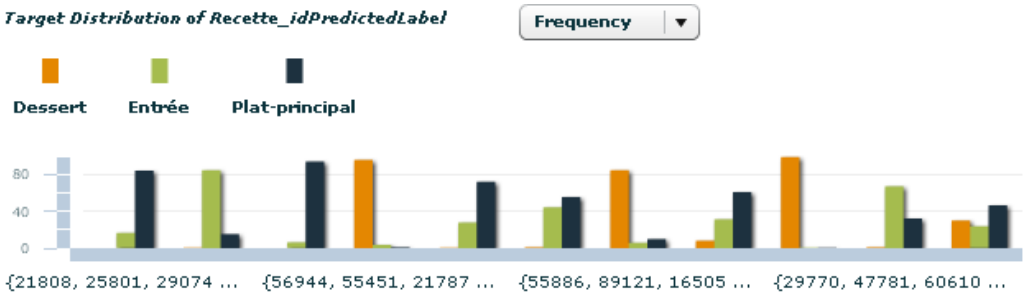


FIGURE 5 – Distribution de clusters

3.2.3 Essai 3

Le 3ème essai reprend le même algorithme que l'essai 2, mais le co-clustering a été réalisé non pas sur les données d'apprentissage mais sur l'ensemble des données d'apprentissage et de test, c'est à dire sur environ 16000 recettes.

4 Tâche 3

Cette tâche a été l'occasion de tester une forme de distance sémantique mot/mot généralisée de manière simple à une distance document/document. Nous nous sommes placés dans un cadre applicatif dans lequel les documents courts (titres) sont en quantité limitée mais gagnent à être ordonnés de manière pertinente.

4.1 Prétraitements

Pour cette tâche, l'espace de représentation des documents a été modifié. Nous avons utilisé notre outil de Traitement Automatique des Langues, TiLT (Heinecke *et al.*, 2008), pour effectuer une lemmatisation associée à un étiquetage syntaxique des documents. Certains traitements ont été réalisés pour tous les types de champs (titre, ingrédients, préparation). Les termes liés aux catégories suivantes été retirés pour tous les types de champs : pronoms, prépositions, déterminants, conjonctions, adverbes, interjections, ainsi que les quantités numériques de tout type. Dans tous les cas, nous avons aussi cumulé, dans un même vecteur, les formes obtenues ainsi que les lemmes correspondants, de manière unitaire, afin d'obtenir une liste de termes uniques.
Nous avons ensuite réalisé des filtrages différenciés en fonction des types de champs (titre, ingrédients, préparation). En ce qui concerne les titres, seules les formes ont été conservées, pas leurs lemmes. Pour les ingrédients, ce sont les formes et les lemmes qui ont été retenus. Les préparations sont aussi représentées par leurs formes et leurs lemmes mais elles ont subi un filtrage supplémentaire puisque nous avons aussi ôté les adjectifs et les verbes. En effet, le rapprochement entre titres et recettes ne fait pas ou peu intervenir les adjectifs et les verbes. Toutefois, nous avons conservé les adjectifs et les verbes des titres et des ingrédients car le contexte de lemmatisation générait des erreurs du type nom/adjectif telle que « saumon » qui aurait été filtré suite à une erreur l'étiquetant comme adjectif. Au final, les ingrédients et les préparations correspondantes sont regroupés en un seul vecteur constitué d'éléments uniques. L'utilisation des lemmes est surtout destinée à étendre la représentation

© ATALA

de la recette afin d'augmenter l'intersection entre le vecteur ingrédients-préparation et les mots des titres. Les multi-mots détectés ne sont pas utilisés mais décomposés en mots simples. Le résultat de ce traitement appliqué, par exemple, à la recette « Risotto de quinoa aux champignons » donne la représentation suivante :

Titre : *risotto quinoa champignons*
Document (ingrédients et préparation) : c*hampignons champignon frais quinoa noix oignon sel poivre poivrer persil morceaux morceau sauteuse sauteur huile eau temps risotto poêle bouillon sauce*

Cet exemple montre un taux de recouvrement des mots du titre avec le contenu de la recette de 100%.

Cette représentation a été testée parmi d'autres et s'est avérée nécessaire et suffisante pour cette tâche. Nous limitons ainsi la taille des vecteurs à comparer ce qui limite le temps de calcul notamment le temps nécessaire pour appliquer la mesure de distance présentée au chapitre suivant.

4.2 Distance sémantique

Nous avons utilisé et adapté la mesure proposée par (Cilibrasi et Vitanyi, 2006). Pour rappel, si x et y sont deux termes, f(x) la fréquence de x , f(y) la fréquence de y, f(x,y) la fréquence de x ET y, et M une valeur de normalisation (nombre de documents), la « distance » est la suivante :

$$D(x,y) = \frac{max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \ min\{\log f(x), \log f(y)\}}$$

Cette mesure vérifie les deux premiers critères d'une distance mathématique (symétrie et séparation) mais pas le critère d'inégalité triangulaire :

D(x,y) = D(y,x) (symétrie)
D(x,y) = 0 ⇔ x=y (séparation)
D(x,z) <= D(x,y)+D(y,z) non vérifié

Cette mesure donne donc une forme de proximité sémantique qui varie entre 0 et 1, 0 indiquant que x est « sémantiquement » proche de y et 1 indiquant que x est « sémantiquement » très éloigné de y. Les logarithmes et la constante de normalisation donnent une valeur qu'il faut plutôt interpréter comme une relation d'ordre : x est plus proche de y que de z.

Cette mesure a été initialement proposée en utilisant les comptages fournis par Google. Pour une utilisation concrète faisant appel à un accès intensif à des centaines de millions de distances, l'utilisation de requêtes vers Google n'est pas possible. Nous avons choisi de réaliser ce calcul à partir des documents de Wikipédia français. Une fois le calcul terminé, nous disposons d'une ressource de « distances sémantiques entre termes » accessible en temps réel.

Ce calcul a été réalisé en 2010 à partir d'une version de Wikipédia datant de fin 2009. Les documents utilisés sont les paragraphes issus naturellement des pages de Wikipédia lors du

© ATALA

filtrage des méta-données de formatage des pages de Wikipédia. Le calcul fréquentiel f(x,y) est donc effectué dans le contexte d'une phrase ou d'un paragraphe. Après traitement, nous avons extrait environ 6 000 000 paragraphes et nous avons constitué une matrice de 193 999 258 distances associées à autant de couples de mots simples. La matrice utilisée ici n'intègre ni les entités nommées, ni les multi-mots.

C'est cette ressource que nous avons voulu tester sur la tâche 3, l'idée étant de calculer une distance globale entre chaque titre et chaque recette par le biais des distances individuelles entre chaque mot des titres et de la recette. Ces distances étant issues d'une ressource initialement indépendante de la tâche mais relativement importante par sa taille et la variété de ses contenus, il est intéressant de pouvoir la tester dans un domaine de spécialité.

Autre point important, ce type de mesure permet d'ordonner la liste des titres en fonction des termes de la recette même si les termes du titre n'apparaissent pas parmi les termes de la recette. Il suffit qu'un terme du titre possède une entrée dans notre matrice en relation avec un mot de la recette pour avoir une distance utilisable.

En ce qui concerne la recette du « Risotto de quinoa aux champignons », nous pouvons ainsi potentiellement récupérer les distances entre les mots du titre (*risotto quinoa champignons)* et tous les mots de la recette (c*hampignons champignon frais quinoa noix oignon sel poivre poivrer persil morceaux morceau sauteuse sauteur huile eau temps risotto poêle bouillon sauce).*

Nous pouvons donc réaliser ce calcul pour chaque couple (titre, recette). Si T est le nombre de termes du titre et R le nombre de termes de la recette, le nombre de distances à considérer est T×R. Nous avons ensuite choisi une métrique globale faisant intervenir les distances unitaires et synthétisant la proximité entre deux documents, en l'occurrence un titre et une recette. Des essais préliminaires ont montré qu'il était plus pertinent de retenir, pour chaque terme d'un titre, uniquement le terme de la recette le plus proche, donc de distance minimum. En ce qui concerne la recette précédente, nous obtenons donc naturellement une distance minimum de 0 pour les termes communs au titre et à la recette :

D_{min}(risotto,risotto) = 0.0
D_{min}(quinoa,quinoa) = 0.0
D_{min}(champignons,champignons) = 0.0

Pour d'autres titres, nous obtenons des valeurs comprises entre 0 et 1 pour chaque terme du titre en relation avec l'élément le plus proche de la recette. Par exemple, pour le titre représenté par (gâteau chocolat noix coco) :

D (gâteau,*noix*) = min{ D (gâteau,*m*)| *m* mot de la recette}
= 0.182777617404
D (chocolat,*noix*) = min{ D (chocolat,*m*)| *m* mot de la recette}
= 0.192853987214
D (noix,*noix*) = min{ D (noix,*m*)| *m* mot de la recette}
= 0.0 (terme commun)
D (coco,*noix*) = min{ D (coco,*m*)| *m* mot de la recette}
= 0.131960109813

© ATALA

Pour le titre représenté par (gratin courgettes pain ail)

D (gratin, *champignons*) = min{ D (gratin,*m*)| *m* mot de la recette}
= 0.200901512394
D (courgettes, *sauteur*) = min{ D (courgettes,*m*)| *m* mot de la recette}
= 0.190081305148
D (pain, *persil*,) = min{ D (pain,*m*)| *m* mot de la recette}
= 0.17466164279
D (ail, *oignon*) = min{ D (ail,*m*)| *m* mot de la recette}
= 0.0799472060861

Ces valeurs sont donc le reflet du calcul issu du corpus de Wikipédia. Si on peut établir facilement le lien (ail, oignon) ou (coco, noix) il est plus difficile d'expliquer la relation (courgettes, sauteur) !

La mesure globale que nous avons ensuite retenue est la moyenne de ces distances minimales. Si T est un titre de N termes et R une recette, la distance globale est :

$$D(T,R)=\frac{1}{N}\ \sum_i^j D_{min}(T_i, R_j)$$

Nous obtenons une mesure globale de proximité qui vaut 0 si les termes du titre sont contenus dans les termes de la recette. Cette mesure tend vers 1 si aucun des termes du titre n'est contenu dans la recette et si de plus ces termes sont « éloignés » des termes de la recette. La moyenne permet de normaliser la mesure par rapport à la longueur du titre. Nous avons bien une mesure qui permet d'ordonner les titres de manière relative dans tous les cas, même si aucun terme du titre n'apparaît dans la recette, ou presque… En effet, si cela est vrai pour la majorité des titres, certains titres dont les termes ne sont jamais apparus dans Wikipédia ne possèdent aucun lien avec d'autres termes. Un exemple est le titre « Tiou Yap ». Dans ce cas la distance à chaque recette est de 1 ce qui reste cohérent.

Les 3 essais portant sur cette tâche sont des variantes de cette mesure :

- le premier essai n'utilise pas la matrice de distance issue de Wikipédia mais fixe à 0 la distance d'un terme du titre apparaissant dans la recette et à 1 s'il n'apparaît pas. C'est une sorte de limite binaire par défaut de la distance utilisée.

- le deuxième essai utilise la métrique décrite calculée à partir de Wikipédia français.

- le troisième essai utilise aussi la métrique décrite mais nécessite deux matrices. La première est toujours celle issue de Wikipédia français. La deuxième a été calculée à partir du corpus d'apprentissage de la tâche. Toutes les recettes (titre, ingrédients, préparation) ont servi à calculer une matrice de distances mot/mot relative à ces données. Au final, nous utilisons une forme de vote qui consiste à choisir la distance minimale issue d'une des deux matrices. Nous faisons l'hypothèse que la matrice calculée sur les recettes des données d'apprentissage est plus spécifique et pertinente que celle de Wikipédia mais moins complète.

Pour des raisons de temps les essais 2 et 3 ont été appliqués sur les 50 premiers résultats de l'essai 1 ce qui correspond plutôt à un ré-ordonnancement du premier essai.

© ATALA

5 Résultats et discussions

5.1 Tâche 1

Pour notre premier essai, utilisant le classifieur de type Maximum d'Entropie, les résultats n'ont pas été satisfaisants. La f-mesure est très basse : 0.086.

Concernant l'essai 2 utilisant Khiops, notre 2ème place (avec une f-mesure de 0.612) semble confirmer les qualités de cet outil. Khiops gère plutôt bien cette tâche que nous estimions difficile compte tenu du type et des quantités relatives des données d'apprentissage. Une technique discriminante appliquée à une sélection de variables effectuée par Khiops aurait peut-être pu nous hisser plus haut !

5.2 Tâche 2

Pour cette tâche, deux constats sont à faire. Tout d'abord, les résultats des deux premiers essais sont assez proches avec des techniques différentes. Nous obtenons une f-mesure de 0.821 pour l'essai 1 et une f-mesure de 0.827 pour l'essai 2.

L'algorithme de type Maximum d' Entropie s'en sort plutôt bien avec le jeu de données d'apprentissage fourni. Par contre, la puissance de modélisation de Khiops ne semble pas avoir permis de surpasser les algorithmes concurrents. Nous sommes un peu surpris du positionnement, compte tenu des résultats habituels de Khiops. Son utilisation sur des variables majoritairement textuelles, ce qui n'est pas son domaine applicatif standard, nécessite peut-être des avancées algorithmiques ou bien des prétraitements amont plus adaptés.

Pour le 3ème essai nous avons réalisé un co-clustering sur les données d'apprentissage et de test. Cette technique améliore généralement, au moins légèrement, les performances d'un co-clustering réalisé uniquement sur les données d'apprentissage (essai 2). Compte tenu des résultats inférieurs du 3ème essai, nous suspectons un problème technique lors de notre essai.

Toujours est-il que Khiops nous donne une vision synthétique des données qui peut être utile dans d'autres cadres applicatifs : sous-catégorisation plus spécialisée des types de plats ou rapprochement de recettes par rapport à peu d'exemples de recettes prototypiques. Il nous a aussi permis d'analyser rapidement le chevauchement des classes *entrée/plat principal* et des variables associées.

5.3 Tâche 3

Pour cette tâche, nous avions choisi de tester une métrique particulière. À la vue de nos résultats, il semblerait que ce choix soit pertinent. Lors de l'évaluation, nous obtenons un MRR (Mean Reciprocal Rank) de 0.4301 pour l'essai 1, de 0.4233 pour l'essai 2 et finalement de 0.4344 pour l'essai 3. Si nous regardons de plus près les résultats des différents essais, ce choix est moins évident. En effet, la distance binaire naïve du 1er essai produit un meilleur résultat que le 2ème essai qui utilise les distances mot/mot générales issues de Wikipédia. Il semblerait donc que la prise en compte de ces valeurs réelles entre 0 et 1 agisse plutôt comme du bruit. On peut aussi penser que notre stratégie qui a consisté à utiliser un pré-

© ATALA

ordonnancement des cinquante premiers titres avec une métrique naïve, puis à réordonner ces titres avec une deuxième métrique élimine des bonnes solutions qui sont en quelque sorte oubliées…

Le 3ème essai qui donne les meilleurs résultats en utilisant l'information spécifique au domaine (corpus d'apprentissage) nous encourage toutefois à continuer dans l'exploration de cette voie.

6 Conclusion

Cette participation à DEFT2013 nous a permis de tester des algorithmes développés par Orange Labs sur des données qui peuvent paraître loin de nos préoccupations habituelles. La spécificité des problèmes de classification des tâches 1 et 2 nous intéressait justement par leur production peu classique. L'outil testé (Khiops) montre ses capacités à s'adapter à de nombreuses tâches et particulièrement ici à des variables majoritairement textuelles. Il a montré ses performances en classification supervisée sur la tâche 1. Il se montre un peu moins performant sur la tâche 2, bien que présentant des résultats tout à fait honorables. Ses capacités intégrées d'analyse non supervisée en font un outil globalement très performant et utile pour de nombreuses tâches. La tâche 3 a été l'occasion d'expérimenter une métrique document/document basée sur une métrique individuelle mot/mot. Ces premiers essais nous paraissent intéressants et demandent à être confirmés sur d'autres types de données, particulièrement sur des documents possédant peu de mots en commun. Finalement, la tâche 3 s'est avérée être presque un cas d'usage que nous allons exploiter dans le cadre du « AI Mashup Challenge ».

Remerciements

Nous remercions Marc Boullé pour ses conseils sur les tâches de classification ainsi que pour son aide avec Khiops.

Références

BELAUNDE, M., PINSON F., COLLIN, O. (2013). Cooking Assistant mashup. *AI Mashup Challenge (ESCW 2013),* Accepted

BOULLÉ, M. (2007). Compression-Based Averaging of Selective Naive Bayes Classifiers. *Journal of Machine Learning Research,* 8:1659-1685.

BOULLÉ, M. (2008). Khiops: outil de préparation et modélisation des données pour la fouille des grandes bases de données. *In Extraction et gestion des connaissances, EGC'2008,* Sophia-Antipolis, France

BOULLÉ, M. (2011). Data grid models for preparation and modeling in supervised learning. *In Hands-On Pattern Recognition: Challenges in Machine Learning, volume 1, I. Guyon, G. Cawley, G. Dror, A. Saffari (eds.),* pages 99-130, Microtome Publishing.

CILIBRASI, R., VITANYI, P. (2006). Similarity of Objects and the Meaning of Words. *In Proceedings of the Third international conference on Theory and Applications of Models of Computation, TAMC'06*, Beijing, China, pages 21-45.

© ATALA

CILIBRASI, R., VITANYI, P. (2007). The Google similarity distance. *In IEEE Transactions on Knowledge and Data Engineering,* 19(3), 370–383 (2007). Preliminary version: "Automatic Meaning Discovery Using Google", Arxiv preprint cs.CL/0412098, 2004, arxiv.org

CILIBRASI, R., BALBACH, F. J., VITANYI, P., LI, M. (2009). Normalized Information Distance *Information Theory and Statistical Learning (*Franck Emmert-Streib, Matthias Dehmer) ISBN: 978-0-387-84815-0 (Print) 978-0-387-84816-7 (Online), pages 45-82, Chapitre 3

HEINECKE, J., SMITS, G., CHARDENON, C., GUIMIER DE NEEF, E., MAILLEBUAU, E., BOUALEM, M. (2008). TiLT : plate-forme pour le traitement automatique des langues naturelles. *Traitement automatique des langues ,* 49(2):17-41.

NIGAM, K., LAFFERTY, J., McCALLUM, A. (1999). Using Maximum Entropy for text classification. *In Proceedings of the IJCAI Workshop on Information Filtering*, pages 61-67

TSURUOKA, Y., TSUJII, J., ANANIADOU, S. (2009). Stochastic Gradient Descent Training for L1-regularized Log-linear Models with Cumulative Penalty. *In Proceedings of ACL-IJCNLP*, pages 477-485

© ATALA

Trois recettes d'apprentissage automatique pour un système d'extraction d'information et de classification de recettes de cuisine *

Eric Charton[1] Ludovic Jean-Louis[1] Marie-Jean Meurs[2] Michel Gagnon[1]
(1) WeST, Ecole Polytechnique de Montréal
(2) Centre for Structural and Functional Genomics, Concordia University
Montréal, QC, Canada
{eric.charton, michel.gagnon, ludovic.jean-louis}@polymtl.ca,
marie-jean.meurs@concordia.ca

RÉSUMÉ

Cet article présente la participation du Wikimeta Lab au Défi Fouille de Texte (DeFT) 2013. En 2013, le défi s'intéresse à la fouille de recettes de cuisine en langue française et se décline en trois tâches de classification et une tâche d'extraction d'information. Les recettes de cuisine sont issues d'un site internet collaboratif et sont donc conçues par des internautes, avec très peu de contraintes quant à la définition des différents paramètres, ingrédients et commentaires qui les composent. Cette particularité rend certaines caractéristiques des corpus difficiles à modéliser dans un système d'apprentissage automatique, faute de régularité dans les données fournies. Nous expliquons dans cette communication notre démarche pour élaborer malgré cette contrainte, plusieurs systèmes ayant obtenus des résultats intéressants dans les tâches 1, 2 et 4.

ABSTRACT

Wikimeta Lab participation in DeFT 2013 - Machine Learning for Information Extraction and Classification of Cooking Recipes

This paper presents Wikimeta Lab participation in the Défi Fouille de Texte (DeFT) 2013. In 2013, this evaluation campaign is focused on mining cooking recipes in French. The campaign consists of three classification tasks and an information extraction task. The corpus is composed of recipes from a collaborative web site, thus they are written by users with almost no constraints on labels, ingredients, and comments. This very context makes some of the corpus specificities difficult to model for a machine learning based system. In this paper, we explain our approach for buiding relevant and effective systems dealing with such a corpus.

MOTS-CLÉS : classification, extraction d'information, DeFT2013, recette de cuisine.

KEYWORDS: classification, information extraction, DeFT2013, cooking recipe.

*. Les auteurs remercient Wikimeta Technologies Inc d'avoir mis ses moyens, son laboratoire et sa cuisine à leur disposition lors de ce défi.

 © ATALA

1 Introduction

Cet article présente les systèmes que nous avons développés pour participer aux tâches 1, 2 et 4 du Défi Fouille de Texte (DeFT) 2013. Cette année, le défi s'intéresse à la fouille de recettes de cuisine en langue française. Le corpus est composé de recettes extraites du site Marmiton.org (`http://www.marmiton.org/`). Le défi propose trois tâches de classification (tâches 1 à 3) et une tâche d'extraction d'information (tâche 4). La tâche 1 consiste à identifier à partir du titre et du texte de la recette son niveau de difficulté sur une échelle à 4 niveaux (très facile, facile, moyennement difficile, difficile). La tâche 2 consiste à identifier le type de plat préparé (entrée, plat principal, dessert) à partir du titre et du texte de la recette. La tâche 3 consiste à apparier le texte d'une recette à son titre. Enfin, la tâche 4 consiste à extraire du titre et du texte d'une recette la liste de ses ingrédients. Notre système développé pour la tâche 1 (détection du niveau de difficulté) est classé premier sur six. Nos systèmes développés pour les tâches 2 (détection du type de plat) et 4 (extraction des ingrédients) sont classés seconds sur cinq.

Cet article est organisé comme suit : La section 2 décrit tout d'abord les corpus utilisés pour les tâches 1, 2 et 4 du défi. Les sections 3, 4 et 5 présentent ensuite les systèmes que nous avons développés pour traiter respectivement la première, la seconde et la quatrième tâche, ainsi que les résultats expérimentaux obtenus par ces systèmes sur le corpus d'entraînement. Enfin, la section 6 conclut cet article et propose des pistes d'investigations complémentaires.

2 Analyse du corpus

Le corpus utilisé pour le défi 2013 est composé de recettes de cuisine extraites du site Marmiton. Marmiton.org est l'un des sites culinaire francophone de large audience avec plus de 300.000 visiteurs par jour (chiffres Smart AdServer mars 2010, source : `http://www.marmiton.org/`). Les recettes rassemblées dans la base de données de Marmiton.org depuis 1999 sont proposées par les internautes via un formulaire validé pour publication par l'équipe de Marmiton. Lors de la soumission d'une recette, les internautes doivent indiquer le type de plat, le niveau de difficulté, le coût et le type de cuisson en sélectionnant des valeurs parmi des listes de choix pré-établies. Les paramètres numériques de la recette tels les temps de préparation et de cuisson, et le nombre de convives, sont à renseigner dans des champs contraints. Les ingrédients, les consignes de préparation et la boisson conseillée sont recueillis dans des champs en texte libre.

Le corpus d'entraînement contient 13.864 recettes pour un volume de données de 19,2 MB. Les corpus de test pour les tâches 1, 2 et 4 sont composés respectivement de 2309, 2307 et 2306 recettes pour des volumes de données de 3MB, 2,9MB et 2,2MB. Les recettes sont fournies au format XML. Chaque fichier du corpus d'entraînement contient le titre de la recette, son type, son niveau, son coût, la liste non normalisée de ses ingrédients et leur quantité d'usage, ainsi que les indications de préparation en texte libre.

2.1 Corpus associé à la tâche 1 : classification par niveau de difficulté

Le tableau 1 détaille le nombre de recettes du corpus d'entraînement selon leur niveau. Les quatre niveaux de difficulté possibles pour chaque recette sont présents de manière très inégale dans le

 © ATALA

corpus d'apprentissage. On constate un fort déséquilibre entre les classes "Très facile" et "Facile" qui contiennent à elles seules plus de 90% des recettes, et les classes "Moyennement difficile" et "Difficile" qui représentent respectivement moins de 10% et 1% du corpus d'apprentissage.

Niveau	corpus d'entraînement		corpus de test	
	#recettes	% du corpus	#recettes	% du corpus
Très facile	6962	50,2	1132	49,0
Facile	5752	41,5	968	41,9
Moyennement difficile	1068	7,7	189	8,2
Difficile	80	0,6	20	0,9
Total	13862*		2309	

* Deux recettes sont incorrectement étiquetées *tres-facile*.

TABLE 1 – Répartition des recettes selon leur niveau de difficulté

2.2 Corpus associé à la tâche 2 : classification par type de plat

Le tableau 2 détaille le nombre de recettes du corpus d'entraînement selon leur type.

La répartition des classes de la tâche 2 dans le corpus d'apprentissage est plus homogène que pour la tâche 1. On constate cependant que presque une recette sur deux du corpus d'apprentissage décrit la réalisation d'un plat principal.

Type de plat	corpus d'entraînement		corpus de test	
	#recettes	% du corpus	#recettes	% du corpus
Entrée	3246	23,4	562	24,4
Plat principal	6449	46,5	1084	47,0
Dessert	4169	30,1	661	28,6
Total	13864		2307	

TABLE 2 – Répartition des recettes selon le type de plat préparé

2.3 Corpus associé à la tâche 4 : extraction des ingrédients

Pour cette tâche, la liste normalisée des ingrédients à extraire contient 960 ingrédients parmi lesquels 102 n'apparaissent pas dans le corpus d'entraînement et 298 n'apparaissent pas dans le corpus de test. Parmi ceux effectivement présents dans le corpus d'entraînement, 348 apparaissent dans 1 à 10 recettes, 355 apparaissent dans 10 à 100 recettes, 138 apparaissent dans 100 à 1000 recettes et enfin, 17 apparaissent dans plus de 1000 recettes.

Le tableau 3 montre la liste des 20 ingrédients les plus fréquents dans les corpus d'entraînement et de test, ainsi que le nombre de recettes dans lesquelles ils sont utilisés. Les données du tableau 3 sont en accord avec celles du tableau 2 et confirment, sans surprise, que les ingrédients prépondérants sont des éléments de base pour l'élaboration de plats principaux et de desserts (*sel, poivre, oignon, sucre, lait, farine, etc.*). On relève aussi que les ingrédients les plus fréquents sur le corpus de test sont identiques à ceux du corpus d'entraînement (les rangs sont quasi similaires).

© ATALA

Rang	corpus d'entraînement			corpus de test		
	ingrédient	# recettes	% du corpus	ingrédient	# recettes	% du corpus
1	sel	6981	50,4	sel	1122	48,7
2	poivre	6109	44,1	poivre	1009	43,8
3	oeuf	5570	40,2	oeuf	895	38,8
4	beurre	4237	30,6	beurre	733	31,8
5	oignon	3452	24,9	farine	577	25,0
6	farine	3249	23,4	oignon	568	24,6
7	huile-d-olive	2631	19,0	huile-d-olive	438	19,0
8	ail	2561	18,5	ail	430	18,6
9	sucre	2533	18,3	sucre	429	18,6
10	lait	2031	14,6	lait	327	14,2
11	tomate	1554	11,2	tomate	256	11,1
12	huile	1299	9,4	citron	236	10,2
13	persil	1267	9,1	huile	210	9,1
14	citron	1253	9,0	eau	202	8,8
15	eau	1236	8,9	persil	202	8,8
16	echalote	1083	7,8	carotte	183	7,9
17	carotte	1073	7,7	echalote	182	7,9
18	pomme-de-terre	961	6,9	pomme-de-terre	150	6,5
19	pomme	893	6,4	pomme	142	6,2
20	sucre-en-poudre	814	5,9	sucre-en-poudre	133	5,8

TABLE 3 – Liste des 20 ingrédients les plus fréquents dans les corpus d'entraînement et de test

2.4 Quelques perles

Le corpus d'apprentissage contient quelques recettes surprenantes, tant par leurs ingrédients que par leurs conditions de réalisation. Ainsi, la recette 17129 propose de cuisiner un *"gigôt bitume"* (plus exactement six) et requiert *"1 chantier de bâtiment ou de travaux publics"* :

```
<recette id="17129">
<titre>Gigot bitume</titre>
<type>Plat principal</type>
<niveau>Difficile</niveau>
<cout>Moyen</cout>
<ingredients>
  <p>6 gigots d'agneau</p>
  <p>fines herbes, dont thym</p>
  <p>sel, poivre</p>
</ingredients>
<preparation>
Matériel : 1 chantier de bâtiment ou de travaux publics
Envelopper les gigots assaisonnés dans plusieurs couches serrées de papier
kraft-aluminium utilisé en bâtiment. Entourer généreusement de fil de fer pour assurer
que l'ensemble ne se défasse pas. Plonger pendant 25 mn les gigots ainsi préparés dans
le baril de bitume brûlant destiné à l'étanchéité de la terrasse du bâtiment ou au
revêtement de la route en construction. Retirer et défaire avec précaution.
Excellent plat traditionnel de BTP mais qui ne relève pas de la cuisine
familiale. Je le donne pour information suite à l'appel aux gourmands. Rouge
</preparation>
</recette>
```

On citera également la recette 10635 du *Gâteau au fromage* dont la réussite impose de *"psalmodier gentiment des incantations magiques"* telles "*abracadabragrandmèraidemoi, oulalajamaisjepourrai-jamaismangertoutçajenaidéjàpleinlesdoigts...*".

© ATALA

3 Tâche 1 - Identification du niveau de difficulté d'une recette

Les systèmes de classification élaborés pour les tâches 1 et 2 sont à base d'apprentissage automatique. Leurs performances sont donc conditionnées par les qualités discriminantes des caractéristiques sélectionnées pour leur entraînement. Pour résoudre les deux tâches de classification de ce défi, nous avons élaboré un ensemble de paramètres discriminants. Ces paramètres peuvent être aussi bien de type classique en fouille de texte (n-grammes, éléments de champs lexicaux), mais aussi plus spécifiques à la tâche, tels les noms d'ingrédients normalisés, le nombre de mots contenus dans une section (le titre, le descriptif de recette) ou encore les quantités d'ingrédients.

3.1 Caractéristiques discriminantes pour la tâche 1

La tâche 1 concerne l'identification d'un niveau de difficulté dans une recette. La particularité de cette tâche est que l'étiquette de difficulté fournie dans le corpus de référence ne provient pas d'un processus d'annotation classique (avec guide fourni aux annotateurs) mais d'un procédé collaboratif : chaque lecteur qui soumet sa recette sur le site Marmiton est laissé entièrement libre quant à la détermination de son degré de difficulté. Ce critère étant hautement subjectif et relié à l'expertise de l'auteur, il en résulte une définition de difficulté extrêmement variable selon les fiches et difficile à modéliser. Les 78 caractéristiques finalement retenues pour l'entraînement des classifieurs pour la tâche 1 sont les suivantes :

- le nombre de mots du titre,
- le nombre de mots des consignes de préparation,
- le nombre d'ingrédients,
- le coût,
- un sous-ensemble de mots discriminants du vocabulaire spécifique de la classe *Moyennement difficile*,
- un sous-ensemble de trigrammes de mots discriminants,
- le nombre de verbes utilisés dans les consignes pour trois familles de verbes.

Mots discriminants du vocabulaire spécifique de la classe *Moyennement difficile*.
Les mots discriminants du vocabulaire spécifique de la classe *Moyennement difficile* sont obtenus en extrayant tout d'abord les mots du titre et des consignes de préparation pour chaque classe. Ces listes sont ensuite filtrées à l'aide d'un anti-dictionnaire permettant de retirer les mots vides. On obtient alors M_{TF}, M_F, M_{MD} et M_D les vocabulaires associés respectivement aux classes *Très facile* (TF), *Facile* (F), *Moyennement difficile* (MD) et *Difficile* (D).

Pour chaque mot m dans M_c où $c \in \{TF, F, MD, D\}$, on calcule les fréquences d'apparition par classe F^m_{TF}, F^m_F, F^m_{MD} et F^m_D.

La classe *Difficile* étant très peu représentée dans le corpus, nous avons choisi de ne pas la considérer et d'orienter notre sélection vers le vocabulaire spécifique de la classe *Moyennement difficile*.

Ainsi, pour chaque mot m, nous avons calculé Δ^m_{MD}, la différence moyenne de fréquence d'apparition entre la classe *Moyennement difficile* et les classes *Très facile* et *Facile* :

© ATALA

$$\Delta_{MD}^{m} = \frac{2F_{MD}^{m} - F_{TF}^{m} - F_{F}^{m}}{2}$$

On obtient alors l'ensemble M_{MD}^{disc}, des mots que nous nommons ici "mots discriminants du vocabulaire spécifique de la classe *Moyennement difficile*", en sélectionnant tous les mots pour lesquels cette différence moyenne est supérieure à 2% :

$$M_{MD}^{disc} = \{m \in M_{MD}, \quad \Delta_{MD}^{m} \geq 2\%\}$$

Enfin, le sous-ensemble $M_{MD}^{car} \subset M_{MD}^{disc}$ est construit en utilisant l'algorithme CFS (Correlation based Feature Selection) (Hall, 1999) associé à un algorithme de recherche glouton. CFS évalue la valeur d'un sous-ensemble de caractéristiques en considérant leurs capacités de prédiction et leurs degrés de redondance.

Trigrammes de mots discriminants.
L'ensemble de tous les trigrammes présents dans les consignes de préparation du corpus d'apprentissage est extrait. Un sous-ensemble de trigrammes discriminants est ensuite construit en utilisant l'algorithme CFS mentionné précédemment.

Familles de verbes.
Les verbes contenus dans les recettes ont été regroupés en deux ensembles : d'une part ceux qui se retrouvent dans une recette *Facile* ou *Très facile* (ensemble TFF), d'autre part ceux qui se retrouvent dans une recette *Difficile* ou *Moyennement difficile* (ensemble MDD). Pour chaque verbe de chacun de ces deux ensembles, on calcule son nombre total d'apparitions dans les recettes correspondant aux deux niveaux de difficulté. Puis, pour chaque verbe de chaque ensemble, on normalise sa fréquence en la divisant par la somme totale de fréquences pour tous les verbes. Le même procédé a été appliqué en ne prenant que l'ensemble des recettes *Très facile* d'une part (ensemble TF), et l'ensemble MDD d'autre part.

Par la suite, trois ensembles de verbes discriminants ont été obtenus selon leur ratio de fréquence normalisée. Pour chaque ratio α on a un ensemble de verbes discriminants $VD(\alpha)$ tel que :

$$VD(\alpha) = \left\{ v \in MDD \cup TFF \;\middle|\; \begin{array}{l} \alpha \times FN_{MDD}(v) \leq FN_{TFF}(v) \leq (1/\alpha) \times FN_{MDD}(v) \\ \vee \\ \alpha \times FN_{MDD}(v) \leq FN_{TF}(v) \leq (1/\alpha) \times FN_{MDD}(v) \end{array} \right\}$$

où FN_i est la fréquence normalisée du verbe v dans l'ensemble i.

En retirant les verbes déjà contenus dans les ensembles extraits pour un α supérieur, on obtient les trois ensembles suivants :

$VD(20) = \{$ tripler, peser $\}$
$VD(14) = \{$ accorder, étirer, manipuler, redéposer, redevenir, tempérer $\}$
$VD(8) = \{$ aérer, braiser, détendre, échapper, effectuer, masquer, renverser, reprendre $\}$

3.2 Système 1T1 pour la tâche 1

Le système 1T1 est basé sur un réseau bayésien (Pearl, 1986, 1998) dont la structure est apprise sur l'ensemble de caractéristiques décrit en 3.1 (78 variables discrètes) en utilisant l'algorithme K2 (Cooper et Herskovits, 1992), et dont la distribution de probabilités est calculée sur le corpus d'apprentissage par un estimateur simple.

 © ATALA

Soit $X_1, ..., X_n$, variables aléatoires discrètes définies par leur loi jointe P. Ces variables sont représentées par les nœuds $v_i \in V$ du graphe orienté $G(V, E)$ associé au réseau bayésien.

Les arcs $e_i \in E$ du graphe associé au réseau bayésien représentent les dépendances entre les variables. On dit que $u \in V$ est un parent de $v \in V$ si $(u, v) \in E$. L'ensemble des nœuds parents d'un nœud v est noté $pa(v)$. Les fils de v sont les nœuds dont v est un parent. Les descendants de v sont les nœuds fils de fils et leurs descendants.

La structure graphique d'un réseau bayésien satisfait le critère de *d-séparation* : toute variable est indépendante de tout sous-ensemble de ses non-descendants, conditionnellement à ses parents. Cette propriéte permet d'écrire la loi jointe sous la forme : $P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | pa(X_i)$ qui définit complètement le réseau bayésien.

Les expériences ont été réalisées en utilisant les classes BayesNet dans Weka (Hall *et al.*, 2009).

3.2.1 Résultats

Sur le corpus d'entraînement, avec une validation croisée de pas 5, le système 1T1 obtient les résultats présentés dans les tableaux 4 et 5.

Classe	Précision	Rappel	F-mesure
Très facile	0,675	0,753	0,712
Facile	0,575	0,533	0,554
Moyennement difficile	0,389	0,248	0,303
Difficile	0,207	0,207	0,207
Moyenne pondérée	0,609	0,620	0,612

TABLE 4 – Résultats obtenus par le système 1T1 sur le corpus d'entraînement pour la tâche 1

	Classe estimée			
Classe réelle	Très facile	Facile	Moyennement difficile	Difficile
Très facile	5244	1652	59	7
Facile	2336	3068	331	17
Moyennement difficile	181	581	265	41
Difficile	8	31	26	17

TABLE 5 – Matrice de confusion du système 1T1 sur le corpus d'entraînement pour la tâche 1

Sur le corpus de test, le système 1T1 a obtenu les résultats présentés dans le tableau 6 :

Macro évaluation			Micro évaluation		
Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
0,460	0,419	0,438	0,609	0,609	0,609

TABLE 6 – Résultats obtenus par le système 1T1 sur le corpus de test pour la tâche 1

© ATALA

3.3 Système 2T1 pour la tâche 1

Le système 2T1 est basé sur l'algorithme Logistic Model Tree (LMT) (Landwehr *et al.*, 2005) qui combine arbres de décision et modèles de régression logistique.

Un LMT est un arbre composé d'une structure d'arbre de décision standard de taille réduite, avec des fonctions de régression logistique (Collins *et al.*, 2002) au niveau des feuilles. Les paramètres des fonctions de régression logistique sont calculés pour maximiser les probabilités sur les données observées. Le LMT est un classifieur probabiliste dont les résultats sont généralement pertinents lorsque l'on dispose de peu de données d'apprentissage. La structure d'arbre permet de minimiser les erreurs d'entraînement tandis que la régression logistique évite le sur-apprentissage en limitant la taille de l'arbre.

Les expériences ont été réalisées en utilisant les classes LMT (Landwehr *et al.*, 2005; Sumner *et al.*, 2005) dans Weka (Hall *et al.*, 2009).

3.3.1 Résultats

Sur le corpus d'entraînement, avec une validation croisée de pas 5, le système 2T1 obtient les résultats présentés dans les tableaux 7 et 8.

Classe	Précision	Rappel	F-mesure
Très facile	0,671	0,777	0,720
Facile	0,587	0,561	0,574
Moyennement difficile	0,549	0,147	0,232
Difficile	0,400	0,073	0,124
Moyenne pondérée	0,625	0,635	0,618

TABLE 7 – Résultats obtenus par le système 2T1 sur le corpus d'entraînement pour la tâche 1

	Classe estimée			
Classe réelle	Très facile	Facile	Moyennement difficile	Difficile
Très facile	5406	1534	22	0
Facile	2446	3228	76	2
Moyennement difficile	197	707	157	7
Difficile	12	33	31	6

TABLE 8 – Matrice de confusion du système 2T1 sur le corpus d'entraînement pour la tâche 1

Sur le corpus de test, le système 2T1 a obtenu les résultats présentés dans le tableau 9 :

Macro évaluation			Micro évaluation		
Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
0,682	0,375	0,484	0,625	0,625	0,625

TABLE 9 – Résultats obtenus par le système 2T1 sur le corpus de test pour la tâche 1

© ATALA

4 Tâche 2 - Identification du type de plat préparé

Le corpus d'entrainement proposé pour la tâche de détection du type de plat est beaucoup plus homogène que celui de détection de difficulté. L'analyse multivariée des distributions de n-grammes ou des noms d'ingrédients laisse apparaître des marqueurs très spécifiques tels que les noms de fruits pour les desserts, ou les noms de viande pour les plats principaux. Nous avons donc décidé de faire reposer notre approche de classification principalement sur des paramètres de type ingrédients (qui représentent 1.200 paramètres sur les 1.287 sélectionnés).

4.1 Caractéristiques discriminantes

Les 1.287 caractéristiques retenues pour l'entraînement du classifieur pour la tâche 2 sont les suivantes :
- le nombre de mots du titre,
- le nombre de mots des consignes de préparation,
- le nombre d'ingrédients,
- le coût,
- les ingrédients normalisés et sélectionnés par analyse en composantes principales (ACP),
- un sous-ensemble de trigrammes de mots discriminants,
- le nombre de verbes utilisés dans les consignes pour trois familles de verbes.

Sélection de marqueurs d'ingrédients par ACP.
Notre approche de sélection d'une liste d'ingrédients consiste à collecter la totalité des noms d'ingrédients fournis dans la section dédiée du corpus, puis à conserver les plus fréquents. Nous collectons 3.860 unités lexicales telles que *daurade* ou *jus de citron*. Puis nous utilisons chacune de ces unités en tant que détecteur binaire à l'aide d'expressions régulières que nous appliquons sur chacun des textes des recettes. Nous obtenons ainsi 13.864 vecteurs (un par recette) de 3.860 paramètres. Nous soumettons ces 13.864 vecteurs à une analyse en composantes principales configurée pour regrouper les paramètres par groupes de 5 et les ordonner par potentiel discriminant. Il est ainsi possible d'identifier 1.232 paramètres fortement discriminants qui sont conservés en tant que paramètres d'ingrédients normalisés et utilisés pour construire le corpus d'entraînement final.

4.2 Système pour la tâche 2

Le système développé pour traîter la tâche 2 est basé sur un Séparateur à Vaste Marge (SVM, Support Vector Machine) (Vapnik, 1995) à noyau linéaire. Ce choix est motivé par l'aptitude des méthodes à base de SVM à traiter des problèmes de grande dimension. Essentiellement dépendantes des vecteurs supports, elles produisent des résultats pertinents même si les données d'apprentissage sont peu nombreuses. Elles offrent ainsi un bon compromis entre capacité de généralisation et complexité.

On dispose d'un ensemble X de n données étiquetées et d'un ensemble fini U de k classes . Dans le contexte de la tâche 2, $n = 13.864$ et $k = 3$. Chaque donnée $x_{i \in [\![1,n]\!]}$ est caractérisée par p caractéristiques et par sa classe $u_i \in U$. Les données sont décrites vectoriellement dans un espace de dimension p. Pour notre système, $p = 1287$, le classifieur est basé sur un SVM multiclasses (3 classes, une par type de plat) avec un l'approche un-contre-un.

Les expériences ont été réalisées en utilisant les classes libSVM (Chang et Lin, 2011; El-Manzalawy et Honavar, 2005) dans Weka (Hall *et al.*, 2009).

© ATALA

4.3 Résultats

Sur le corpus d'entraînement, avec une validation croisée de pas 5, le système T2 obtient les résultats présentés dans les tableaux 10 et 11.

Classe	Précision	Rappel	F-mesure
Plat principal	0,834	0,854	0,844
Dessert	0,967	0,982	0,974
Entrée	0,694	0,648	0,670
Moyenne pondérée	0,841	0,844	0,842

TABLE 10 – Résultats obtenus par le système T2 sur le corpus d'entraînement pour la tâche 2

	Classe estimée		
Classe réelle	Plat principal	Dessert	Entrée
Plat principal	5507	60	882
Dessert	29	4094	46
Entrée	1064	80	2102

TABLE 11 – Matrice de confusion du système T2 sur le corpus d'entraînement pour la tâche 2

Sur le corpus de test, le système T2 a obtenu les résultats présentés dans le tableau 12 :

Macro évaluation			Micro évaluation		
Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
0,850	0,843	0,847	0,856	0,856	0,856

TABLE 12 – Résultats obtenus par le système T2 sur le corpus de test pour la tâche 2

5 Tâche 4 - Extraction des ingrédients

Parmi les ingrédients à extraire, certains ne se retrouvent pas explicitement dans les consignes de préparation : ils peuvent être remplacés par des formes verbales (*saler*, au lieu de *sel*), nominales (*légume* pour *carotte*, ou *viande* pour *escalopes de poulet*), ou encore omis par les utilisateurs. Dans quelques rares cas, les préparations sont très courtes et ne mentionnent pas d'ingrédients : *"Mélangez tous les ingrédients et faire chauffer dans une poêle à crêpes."* (recette *20827*). A l'opposé certains ingrédients présents dans la description peuvent ne pas faire partie de la recette, par exemple lorsqu'il s'agit de suggestions d'accompagnement pour le plat ou d'une proposition alternative (*"glace à la vanille ou une crème anglaise"*).

En analysant les recettes du corpus d'entraînement et la liste des ingrédients à extraire, nous avons mesuré que dans 39,8% des cas un ingrédient à extraire n'était pas mentionné explicitement dans la préparation. Ce chiffre est une moyenne sur tous les ingrédients, et varie selon les ingrédients.

Notre méthode pour l'extraction des ingrédients s'appuie sur le titre et la description de chaque recette et se compose de trois étapes : (i) la normalisation des recettes, (ii) la détection des ingrédients, (iii) la sélection des ingrédients.

© ATALA

Nous avons soumis deux jeux de résultats pour cette tâche 1T4 et 2T4. Les deux premières étapes sont identiques pour les deux soumissions ; seule la phase de sélection des ingrédients diffère. Nous détaillons ci-après chacune des étapes et les soumissions.

5.1 Normalisation des recettes

La phase de normalisation vise à faire une analyse linguistique du contenu textuel des recettes. Précisément, on utilise le lemmatiseur du TreeTagger [1] pour extraire les lemmes de tous les termes d'une recette. Les lemmes sont ensuite désaccentués.

5.2 Détection des ingrédients

La phase de détection vise à identifier tous les ingrédients de la liste normalisée qui sont contenus dans une recette. En pratique, on utilise des expressions régulières pour rechercher chaque ingrédient de cette liste dans les versions normalisées des recettes (sections titre et préparation). En complément, nous avons ajouté des expressions régulières pour substituer à certaines formes verbales ou nominales les ingrédients correspondants : par exemple, "beurrée|beurrez|beurrer" sont remplacés par "beurre". Au total, la détection s'appuie sur une trentaine d'expressions régulières.

5.3 Sélection des ingrédients

La phase de sélection vise à choisir parmi tous les ingrédients détectés dans la phase précédente ceux effectivement retenus pour la recette. Nous avons proposé deux méthodes pour cette sélection. La première, utilisée par le système 1T4, est fondée sur deux critères : la fréquence et la position de la première occurrence de chaque ingrédient dans la recette. Précisément, les ingrédients sont triés par ordre décroissant de fréquence et ordre croissant d'apparition dans la recette.

La seconde méthode, utilisée par le système 2T4, est à base d'apprentissage statistique. La décision d'attribuer ou non un ingrédient à une recette est prise par un classifieur. Le classifieur s'appuie sur les caractéristiques suivantes pour prendre sa décision :
- la présence/absence de l'ingrédient dans le titre de la recette
- la forme normalisée de l'ingrédient
- le nombre d'occurrences de l'ingrédient dans la recette
- la position de la première occurrence de l'ingrédient dans le document : pos_{first}
- la position de la dernière occurrence dans le document : pos_{last}
- le taux de recouvrement du document : $rec_{last} = (pos_{last} - pos_{first})/taille(recette)$
- la profondeur : $prof = (1 - pos_{first})/taille(recette)$

Pour entraîner le modèle de sélection des ingrédients, nous avons testé plusieurs algorithmes d'apprentissage proposés dans l'outil Weka (Hall *et al.*, 2009) : arbres de décisions, Naïve Bayes, BayesNet et Meta Bagging. Les meilleurs résultats ont été obtenus avec les arbres de décisions et sont reportés dans le tableau 13.

1. http ://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

© ATALA

Classe	Précision	Rappel	F-Mesure
Sélectionné	0,806	0,846	0,825
Non sélectionné	0,806	0,759	0,782
Moyenne pondérée	0,806	0,806	0,806

TABLE 13 – Résultats de l'entraînement du modèle de sélection des ingrédients de la tâche 4

Nous reportons dans le tableau 14, les résultats globaux de l'extraction d'ingrédients, c'est à dire en appliquant nos deux approches de sélection d'ingrédients sur les corpus d'entraînement et de test. Les résultats sont reportés en terme de MAP (Mean Average Precision) et obtenus à partir du script d'évaluation officiel.

Système	corpus d'entraînement	corpus de test
1T4	0,5659	0,5678
2T4	0,6702	0,6462

TABLE 14 – Résultats sur l'extraction des ingrédients pour les deux approches proposées (MAP)

6 Conclusion

Nous avons présenté dans cet article trois systèmes dédiés à la résolution des tâches 1, 2 et 4 du Défi Fouille de Texte (DeFT) 2013. La tâche de fouille de texte proposée a été résolue par des approches par apprentissage automatique, impliquant néanmoins une recherche et une sélection poussée des caractéristiques. La particularité de la tâche 1, dont l'étiquette de difficulté était apposée de manière collaborative et non normalisée, rend l'apprentissage délicat. Nous avons tenté de résoudre ce problème en utilisant une méthode de classification originale qui repose sur l'algorithme Logistic Model Tree (LMT) combinant arbres de décision et modèles de régression logistique. Cette méthode a fourni des résultats intéressants. A l'issue de ce défi, l'utilisation des bases documentaires collaboratives comme ressource d'apprentissage, ainsi que la recherche d'algorithmes de classification appropriés pour traiter leurs données, sont des questions qui nous semblent appeler de plus amples investigations. Les systèmes que nous avons développés pour cette campagne, conçus en langage Java et exploitant la plateforme Weka, sont disponibles en version source libre sur `http://www.wikimeta.org/deft2013`. Ils permettent de reproduire la totalité des expériences de classification décrites dans cette communication.

Références

CHANG, C.-C. et LIN, C.-J. (2011). Libsvm : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

COLLINS, M., SCHAPIRE, R. E. et SINGER, Y. (2002). Logistic regression, adaboost and bregman distances. *Machine Learning*, 48(1-3):253–285.

COOPER, G. F. et HERSKOVITS, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347.

 © ATALA

EL-MANZALAWY, Y. et HONAVAR, V. (2005). Wlsvm : Integrating libsvm into weka environment. *Software available at http ://www. cs. iastate. edu/yasser/wlsvm*.

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P. et WITTEN, I. H. (2009). The weka data mining software : an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

HALL, M. A. (1999). *Correlation-based feature selection for machine learning*. Thèse de doctorat, The University of Waikato.

LANDWEHR, N., HALL, M. et FRANK, E. (2005). Logistic model trees. *Machine Learning*, 59(1-2):161–205.

PEARL, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3):241–288.

PEARL, J. (1998). *Bayesian networks*. MIT Press, Cambridge, MA, USA.

SUMNER, M., FRANK, E. et HALL, M. (2005). Speeding up logistic model tree induction. *In Knowledge Discovery in Databases : PKDD 2005*, pages 675–683. Springer.

VAPNIK, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.

© ATALA

Clustering pour grammaires catégorielles du second ordre

Noémie-Fleur Sandillon-Rezer
CNRS, Esplanade des Arts et Métiers, 33402 Talence
LaBRI, 351 Cours de la Libération, 33405 Talence
nfsr@labri.fr

RÉSUMÉ

Le but de nos travaux actuels est de créer un nouvel algorithme d'inférence grammaticale se servant de clusters hiérarchiques de mots pour guider le processus et extraire une grammaire AB (Lambek, 1958). Pour ce faire, nous utilisons deux entrées : des clusters ainsi que des arbres de dérivation partiellement typés. Nous nous focaliserons donc à la fois sur le descriptif de la méthode et sur la gestion des adverbes, qui forment une classe de mots particulièrement complexe à traiter. Nous présenterons des résultats quant à l'efficacité de la méthode en général, calculée en fonction d'un lexique qui représente notre lexique de référence. Nous avons actuellement une équivalence de 91,3% entre les deux lexiques. Nous ferons aussi le point sur notre méthode pour traiter les adverbes.

ABSTRACT

Adverbs and modifiers management when using grammatical inference guided by clustering

The aim of our work is to create a new grammatical inference algorithm, using hierarchical word clusters to guide the process and extract an AB grammar (Lambek, 1958). We will focus on the method description and the adverb management. The results on the general method are computed by comparing the lexicon obtained with our gold lexicon : the equivalence between them is around 91.3 %. We will also describe the results over the adverbs.

MOTS-CLÉS : grammaires catégorielles, clustering hiérarchique, inférence grammaticale, grammaire AB.

KEYWORDS: categorial grammars, hierarchical clustering, grammatical inference, AB grammar.

© ATALA

1 Introduction

Les clusters hiérarchiques permettent de regrouper des éléments en fonction de critères de similarité, que ce soit en fonction de leur contexte ou d'autres valeurs qui leur sont liées. L'idée d'en utiliser pour apprendre une CFG avait déjà été exploitée par Adriaans (1999) à partir de phrases sans structure et par Bernardi *et al.* (2004) pour tagger les mots d'un corpus italien, le citant aussi comme possibilité pour l'inférence grammaticale.
Notre idée est d'utiliser du clustering pour avoir une notion de contexte commun entre les mots qui nous guide lors de l'induction. L'induction en elle même sera faite en suivant une méthode telle qu'utilisée par Buszkowski et Penn (1990) ou Kanazawa (1998), dans un style à la Gold (1967), en utilisant en entrée des arbres ne contenant pas uniquement des variables, mais dont certains types sont fixés après apprentissage utilisant des structures complètement spécifiées (Hockenmaier, 2003; Sandillon-Rezer et Moot, 2011).
La dernière méthode utilise un transducteur d'arbres (Comon *et al.*, 2007) généralisé pour extraire des corpus de Paris VII (Abeillé *et al.*, 2003) et Sequoia (Candito et Seddah, 2012) des arbres de dérivation d'une grammaire AB (voir figure 1) ainsi que les lexiques correspondants.

$$\frac{A/B \quad B}{A}\ [/E] \qquad \frac{B \quad B\backslash A}{A}\ [\backslash E] \qquad \frac{\forall X.A}{A[B/X]}\ [\forall\varepsilon]$$

FIGURE 1 – Règles que nous utilisons pour les grammaires catégorielles

2 Méthodologie générale

Nous sommes partis du constat que les algorithmes usuels d'unification ne prennent pas ou peu en compte le contexte des mots et pallions à ceci en utilisant des clusters : plutôt que d'essayer d'unifier les types d'un même mot, nous allons unifier les types au sein d'un même cluster. En effet, les algorithmes symboliques exhaustifs sont trop complexes, aussi nous avons préféré y substituer une combinaison statistique et symbolique qui fait baisser la complexité. Pour plus de détails, nous conseillons l'article (Sandillon-Rezer, 2013).

Nous extrayons le contexte à partir des arbres syntaxiques des corpus, en créant des "vecteurs" contenant plusieurs informations contextuelles, comme le POS-tag du mot et celui de ses voisins. Ces vecteurs sont ensuite utilisés dans R (Ihaka et Gentleman, 1993) via une méthode de clustering hiérarchique (Ward, 1963) avec la métrique de Manhattan.

Les arbres de dérivation que nous utilisons en entrée ont été binarisés en utilisant notre transducteur. Cependant, une partie des données ont été effacées et remplacées par des variables : en effet, nous ne gardons que les types les plus fréquents (au dessus de 75%) pour les noeuds arguments et le reste est remplacé.

Dans l'ordre, nous unifions en premier lieu les variables des plus petits clusters, puis les cas ne proposant qu'un seul choix par cluster. Lorsqu'il y a plusieurs choix, nous prenons le candidat le plus simple (calculé en fonction du nombre de \ ou / dans le type) ou bien un candidat au hasard.

Pour l'évaluation, nous avons utilisé des extraits de Sequoia et Paris VII. Elle a été effectuée autant sur le ratio de variables unifiées (0.2% pour Sequoia, 0% pour Paris VII) que sur la ressemblance entre les lexiques résultant de l'unification et ceux extraits directement des arbres après transduction. Il y a exactement 81,7% de types parfaitement identiques entre les deux

 © ATALA

lexiques. Les 18,6% qui diffèrent sont découpés en 1,6% d'erreurs et 17% de types qui soit sont présents pour le mot en question mais pas avec la même occurrence, présentant des différences mineures ou bien équivalents au sens du *CCG Bank* (Hockenmaier et Steedman, 2007).

3 Traitement des modificateurs et conjonctions

Une classe de mots qui contient un grand nombre d'erreurs est celle des adverbes : sur un lexique de 103 types donnés à des adverbes choisis aléatoirement, il n'y a que 34,9% de similitude.Une majorité d'adverbes, hélas, ne sont pas de la forme a/a ou $a\backslash a$, avec a un type quelconque. Pour réduire ce pourcentage d'erreurs, nous proposons deux solutions.

Durant la phase d'extraction des arbres de dérivation, nous pouvons appliquer un traitement spécial sur les adverbes, comme montré en figure 2. Ainsi, nous sommes certains que les adverbes auront un type a/a ou $a\backslash a$. Cette étape lisse totalement le problème majeur des adverbes.

En utilisant cette solution, nous réduisons de moitié le nombre de variables présentes avant unification, en passant de 686 variables à 344. La différence entre les deux lexiques tombe alors à 12,8% (soit 732 paires mot-type), avec 1,1% d'erreurs (63 paires).

L'autre solution consiste à employer, comme proposé par (Capelletti et Tamburini, 2009), un type X qui peut être remplacé par n'importe quel autre type du lexique. Par exemple, lorsqu'un noeud $PP-MOD$ est utilisé en fin de phrase, la préposition aura le type $(s\backslash s)/np$, tandis qu'utilisé en tant que frère à droite d'un groupe verbal, il aura par exemple le type $((np\backslash s)\backslash(np\backslash s))/np$ S'il y a besoin d'unifier ces deux types, ils seraient alors tous les deux remplacés par : $\forall X.(X\backslash X)/np$.

Nous n'avons pas encore implémenté cette méthode, cependant, sur un extrait de 80 adverbes, cette solution lisserait 56,3% des erreurs.

Il faut noter que ce traitement particulier pourrait aussi s'appliquer aux conjonctions de coordinations, qui auraient alors comme type $\forall X.(X\backslash X)/X$.

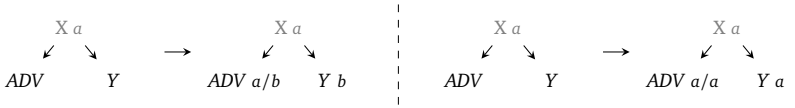


FIGURE 2 – Evolution du traitement des modificateurs : à gauche le traitement initial, à droite celui qui rend transparent la présence des adverbes dans la phrase.

4 Perspectives

Nous voulons tendre vers un meilleur ratio d'équivalence du lexique : le travail commencé avec les adverbes aide grandement, car ils représentent environ 8,5% des mots présents dans notre échantillon de Paris VII et contiennent un grand nombre d'erreurs. Nous comptons aussi tester d'autres méthodes de clustering, comme les *k-means* ou bien le *Clustering By Committee* (Pantel, 2003). Enfin, en créant un cluster avec des phrases venant des différents corpus et d'autres annotées avec un tagger (Moot, 2011), il nous serait possible de typer des phrases sans structure syntaxique en baissant le nombre de dimensions des vecteurs.

© ATALA

Références

ABEILLÉ, A., CLÉMENT, L. et TOUSSENEL, F. (2003). Building a treebank for french. *Treebanks, Kluwer, Dordrecht*.

ADRIAANS, P. W. (1999). Learning shallow Context-Free languages under simple distributions.

BERNARDI, R., BOLOGNESI, A., TAMBURINI, F. et MOORTGAT, M. (2004). Categorial type logic meets dependency grammar to annotate an italian corpus. *In* KRUIJFF, G.-J. M. et DUCHIER, D., éditeurs : *COLING 2004 Recent Advances in Dependency Grammar*, pages 49–56, Geneva, Switzerland. COLING.

BUSZKOWSKI, W. et PENN, G. (1990). Categorial grammars determined from linguistic data by unification. *Studia Logica*.

CANDITO, M. et SEDDAH, D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical.

CAPELLETTI, M. et TAMBURINI, F. (2009). Parsing with polymorphic categorial grammars. *CICLing*.

COMON, H., DAUCHET, M., GILLERON, R., LÖDING, C., JACQUEMARD, F., LUGIEZ, D., TISON, S. et TOMMASI, M. (2007). Tree automata techniques and applications. Available on : `http://www.grappa.univ-lille3.fr/tata`.

GOLD, E. (1967). Language identification in the limit. *Information and Control*, 10.

HOCKENMAIER, J. (2003). Data and models for statistical parsing with combinatory categorial grammar.

HOCKENMAIER, J. et STEEDMAN, M. (2007). CCGbank : a corpus of CCG derivations and dependency structures extracted from the penn treebank. *Computational Linguistics*, 33(3):355–396.

IHAKA, R. et GENTLEMAN, R. (1993). R project.

KANAZAWA, M. (1998). *Learnable Classes of Categorial Grammars.* Center for the Study of Language and Information, Stanford University.

LAMBEK, J. (1958). The mathematics of sentence structure. *The American Mathematical Monthly*, 65.

MOOT, R. (2011). Wide-coverage semantics for spatio-temporal reasoning. *Traitement Automatique des Langues 52*.

PANTEL, P. (2003). Clustering by committee. *PhD thesis*.

SANDILLON-REZER, N.-F. (2013). Inférence grammaticale guidée par clustering. *Recital*.

SANDILLON-REZER, N.-F. et MOOT, R. (2011). Using tree tranducers for grammatical inference. *Proceedings of Logical Aspects of Computational Linguistics 2011*.

WARD, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*.

© ATALA

Des bigrammes au secours de la grammaire

Eric Wehrli, Luka Nerima, Meghdad Farahmand
Laboratoire d'Analyse et de Technologie du Langage
CUI - Université de Genève
Battelle - 7 route de Drize, CH-1227 Carouge
{Eric.Wehrli, Luka.Nerima, Meghdad.Farahmand}@unige.ch

RÉSUMÉ

Ce papier présente les premiers pas d'un projet, qui vise à améliorer la qualité et l'efficacité de l'analyseur syntaxique Fips par l'ajout d'une composante statistique. La technique utilisée repose sur des bigrammes de paires (mot+catégorie) calculés à partir des homographes contenus dans nos lexiques, ainsi que d'un gros corpus de textes extraits de Wikipedia étiqueté au moyen du Stanford Parser. La table de bigrammes doit servir à ordonner les alternatives en cours de traitement de manière à ce que l'élagage épargne les plus probables.

ABSTRACT

Bigrams to help grammar

This paper discusses an on-going project aiming at improving the quality and the efficiency of the Fips syntactic parser by the addition of a statistical component. The proposed technique relies on bigrams of pairs (word+category) selected from the homographs contained in our lexical database and computed over a large corpus from Wikipedia, tagged by means of the Stanford Parser. The bigram table will be used by Fips to rank and prune the set of alternatives during the parsing process.

MOTS-CLÉS : systèmes hybrides, analyse syntaxique, bigrammes.

KEYWORDS: hybrid systems, parsing, bigrams.

1 Introduction

"Peut-on améliorer les performances d'un analyseur syntaxique à base de règles avec des données empiriques ?" c'est la question que nous souhaitons aborder dans ce papier, question qui n'est pas nouvelle, puisqu'elle était abordée il y a près d'une vingtaine d'années déjà (Klavans et Resnik, 1996) et a donné suite à de nombreux travaux. Ces travaux, dans leur grande majorité, ont porté sur l'intégration de données empiriques dans un prétraitement du corpus à analyser, par exemple par l'utilisation d'un segmenteur (*tokenizer*) statistique (Adolphs et al., 2008). La recherche présentée ici, aborde la question de l'hybridation par un autre biais, puisqu'elle cherche à intégrer au sein de l'analyseur syntaxique symbolique Fips (Wehrli, 2007) une composante statistique utilisée pour le filtrage des analyses concurrentes. La technique utilisée repose sur des bigrammes de paires (mot+catégorie) calculés à partir des homographes contenus dans nos lexiques, ainsi que d'un gros corpus de textes extraits de Wikipedia et tagués au moyen du Stanford Parser (Klein et Manning, 2003). La table de

 © ATALA

bigrammes est utilisée par Fips pour ordonner et filtrer les alternatives.

2 Motivations

Un des problèmes auquels font face les analyseurs syntaxiques profonds basés sur des règles et/ou connaissances linguistiques est celui de la croissance exponentielle du nombre d'alternatives à considérer au fur et à mesure du traitement. L'origine du problème est bien connue, il s'agit, en bref, de la conjonction de l'ambiguité lexicale et du nombre de règles susceptibles de s'appliquer dans un contexte donné. La solution habituellement utilisée pour juguler la croissance des alternatives et garantir l'efficacité de l'analyseur est l'élagage ("pruning"), c'est-à-dire l'élimination des alternatives considérées comme les moins probables. Cet élagage intervient lorsque le nombre d'analyses dépasse la largeur du faisceau défini par le nombre maximal d'alternatives spécifié par l'usager. Dans l'analyseur Fips, un score est attribué à chaque alternative, basé sur le type de règle utilisé. Après le traitement de chaque nouveau mot, l'ensemble des alternatives est filtré sur la base de ce score. La figure 1, ci-dessous, montre les résultats obtenus en fonction du nombre maximal d'alternatives sélectionné (colonne 1) sur la base d'un corpus de 52 articles extraits du magazine *The Economist*, totalisant environ 75'000 mots. Le temps de traitement est donné en colonne 2 (minutes :secondes). La colonne 3 exprime le pourcentage d'analyses complètes (c-à-d., une structure couvrant toute la phrase). Enfin, la dernière colonne donne le nombre de mots traités par seconde.

max. alternatives	temps	analyses complètes	mots/secondes
120	21 :06	77.98%	59
80	13 :28	75.45%	93
40	6 :26	71.80%	198
10	2 :17	59.40%	579
5	1 :39	50.94%	793

FIGURE 1 – Résultats obtenus en fonction du nombre d'alternatives

Ces résultats montrent clairement que le nombre d'analyses complètes augmente en fonction du nombre d'alternatives. Sachant que le taux d'analyses complètes est en corrélation avec la qualité des analyses, on constate que l'augmentation du nombre d'alternatives améliore la qualité des résultats. Cette amélioration a un coût, puisque pour gagner moins de 4% d'analyses complètes supplémentaires on double le temps de traitement (cf. lignes 2 et 3 de la figure 1). Comme le choix du nombre d'alternatives n'a aucun impact sur la grammaire, la conclusion est claire, la diminution du nombre d'analyses complètes liée à la réduction du nombre d'alternatives provient du filtrage. Autrement dit, l'élagage de l'ensemble des alternatives élimine des candidats qu'il aurait fallu conserver.

L'objectif de cette recherche consiste donc à améliorer le filtrage, plus précisément à éviter que le filtrage n'élimine de bons candidats. Le résultat attendu de l'intégration du module de bigrammes dans Fips est de diminuer la largeur du faisceau d'analyses, et par conséquent d'accéler le traitement, sans diminuer la qualité des analyses. Par exemple, en reprenant les chiffres de la figure 1, il s'agirait de réduire le nombre maximal d'alternatives de 120 à 40 tout en conservant 77.98% d'analyses complètes.

© ATALA

Lorsque l'analyseur rencontre un homographe [1] la table des bigrammes est consultée pour déterminer la probabilité de chacune des lectures (catégories) du mot homographe étant donné le mot précédent et sa catégorie dans la structure syntaxique déjà construite. Ainsi, dans la phrase (1), lorsque l'analyseur lit le mot homographe *costs* (substantif ou verbe), il consulte la table des bigrammes pour chacune des deux catégories (N et V) étant donné le verbe *labour*, dans une première analyse, le substantif *labour* dans une deuxième analyse. La probabilité associée au bigramme [labour-N, costs-N] (*coût de la main-d'oeuvre*) est largement supérieure à celle des autres bigrammes. Cette analyse sera donc clairement privilégiée. Cette approche se distingue de celle qui consisterait à utiliser un étiqueteur (POS-tagger) en prétraitement, par le fait que les lectures non privilégies sont conservées, jusqu'à un élagage éventuel de l'analyse.

(1) Unit labour **costs** measure the average cost of labour per unit of output.
les coûts unitaires de la main-d'oeuvre mesurent le coût moyen de la main-d'oeuvre par unité produite.

3 Extraction des bigrammes

Sur la base de notre base de données lexicale, qui contient toutes les formes fléchies, nous avons établi la liste des formes homographiques, soit environ 10'000 formes. La table des bigrammes est constituée de tous les couples `e1-e2`, où `e2` est une forme homographique avec sa catégorie lexicale et `e1` le mot qui la précède immédiatement dans le texte, également avec sa catégorie. Chaque bigramme donne la probabilité de `e2`, étant donné `e1`. Ces bigrammes ont été calculés sur la base d'un extrait du Wikipedia anglais d'environ 140 mio de mots, préalablement étiqueté au moyen du Stanford Parser. Près de 7'000 de nos formes homographiques ont été répertoriées dans ce corpus, totalisant 37 mio d'occurrences, soit 26% du corpus total.

Références

Adolphs, P., S. Oepen, U. Callmeier, B. Crysmann, D. Flickinger et B. Kiefer, 2008. "Some Fine Points of Hybrid Natural Language Parsing", *Proceedings of LREC-2008*, Marrakesh, Morocco.

Klavans, J. et Ph. Resnik (éd.), 1996. *The Balancing Act : Combining Symbolic and Statistical Approaches to Language*, MIT Press.

Klein, D. et Ch.D. Manning. 2003. "Accurate Unlexicalized Parsing", *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

Sennrich, R., G. Schneider, M. Volk et M. Warin, 2009. "A New Hybrid Dependency Parser for German", *Proceedings of GSCL-Conference*, Potsdam.

Wehrli, E. 2007. "Fips, a 'deep' linguistic multilingual parser" in *Proceedings of the ACL 2007 Workshop on Deep Linguistic processing*, pp. 120-127, Prague, Czech Republic.

1. Par homographe, il faut entendre ici, mot orthographique qui appartient à plus d'une catégorie lexicale, comme *président*(substantif/verbe), *pas* (adverbe/ substantif), *intime* (adjectif/substantif/verbe), etc.

 © ATALA

Intégrer des ressources lexicales et grammaticales externes dans des analyseurs partiels probabilistes

Matthieu Constant[1] Isabelle Tellier[2]
(1) Université Paris-Est, LIGM, CNRS
(2) Université Paris 3, Lattice, CNRS
mconstan@univ-mlv.fr, isabelle.tellier@univ-paris3.fr

RÉSUMÉ

Cet article traite de segmentation lexicale et de segmentation syntaxique. Nous montrons comment combiner des modèles probabilistes comme les champs markoviens aléatoires avec des ressources lexicales et grammaticales externes à l'aide d'une approche à états finis.

ABSTRACT

Integration of external linguistic resources in lexical and syntactic segmentation

This article focuses on lexical and syntactic segmentation. We show how to combine probabilistic models like conditional random fields with external lexical and grammatical resources using a finite-state approach.

MOTS-CLÉS : Segmentation lexicale, segmentation syntaxique, champs markoviens aléatoires, ressources lexicales et grammaticales.

KEYWORDS: Lexical segmentation, Syntactic segmentation, Conditional Random Fields, Lexical and grammatical resources.

© ATALA

1 Problématique

Dans cet article, nous nous focalisons sur la segmentation lexicale et syntaxique, qui correspond à l'étiquetage grammatical et l'analyse en constituants simples, incorporant la reconnaissance des mots composés. En particulier, nous décrivons une approche pour combiner des modèles probabilistes de segmentation et des ressources lexicales et grammaticales.

Une tâche de segmentation peut revenir à une tâche d'étiquetage simple en ajoutant des informations de segmentation dans les étiquettes (ex. BIO (Ramshaw et Marcus, 1995)). Par exemple, la phrase *Luc boit de l'eau de vie* peut s'étiqueter grammaticalement de la manière suivante *Paul/B-NPP boit/B-V de/B-DET l'/I-DET eau/B-NC de/I-NC vie/I-NC* où *de l'* est un déterminant composé partitif et *eau de vie* est un nom composé. Un analyseur probabiliste cherche à trouver la séquence d'étiquettes la plus probable étant donnée une séquence de même taille de mots (graphiques) en entrée. Les analyseurs *état-de-l'art* sont, en général, basés sur des modèles discriminants tels que maximum d'entropie ou les champs markoviens aléatoires (CRF).

Ces approches probabilistes dépendant uniquement d'un corpus d'apprentissage montrent cependant quelques limites. Par exemple, le traitement des mots simples et composés inconnus [1] cause de sérieuses chutes de performances des analyseurs. Des travaux récents se consacrent à l'utilisation de ressources lexicales riches comme sources de traits des modèles probabilistes afin de résorber en partie le problème : ex. (Denis et Sagot, 2009) pour un étiquetage simple, (Constant et Tellier, 2012) pour la reconnaissance des mots composés. C'est le genre de techniques que nous prolongeons ici.

2 Incorporation de ressources externes sous forme de traits

Nous utilisons les ressources externes comme sources de traits. Pour cela, nous nous basons sur une architecture à états finis (Blanc *et al.*, 2007). Celle-ci comporte d'abord une analyse ambigue du texte par consultation de dictionnaires électroniques et application d'une cascade de transducteurs. Cette procédure génère un automate acyclique représentant l'ensemble des analyses possibles. En particulier, l'analyse d'une séquence figée ou d'un constituant est représentée par une seule transition. Par exemple, l'analyse figée de *eau de vie* sera placée dans une seule transition étiquetée comme un nom. L'automate peut ensuite être élagué itérativement au moyen de différents modules de levée d'ambiguité basés sur des règles ou des heuristiques. A chacune de ces itérations, il est possible d'extraire des traits à partir de l'automate à chacune des positions dans le texte. Pour cela, on convertit l'automate au format BIO. Chaque transition correspondant à une séquence de m mots est convertie en une séquence de m transitions qui attribue une étiquette avec une information de type BIO à chacun des mots. Par exemple, la transition pour le mot composé *eau de vie* étiqueté comme un nom (NC) est décomposé en une séquence de 3 transitions : *eau/B-NC de/I-NC vie/I-NC*. La figure 1 donne l'automate issu de l'analyse lexicale dans le schéma BIO. Ainsi, à chaque position, on peut associer un ensemble de transitions duquel on peut extraire un certain nombre d'informations que l'on peut ajouter sous forme de traits au modèle. Par exemple, pour la position correspondant aux transitions sortant de l'état 5 (le mot *de*), on a l'ensemble d'étiquettes {I-NC,B-DET,B-PREP}. On peut en

1. qui ne se trouvent pas dans le corpus d'apprentissage.

© ATALA

particulier utiliser la concaténation des étiquettes de l'ensemble dans un certain ordre (ex. ordre alphabétique) comme trait.

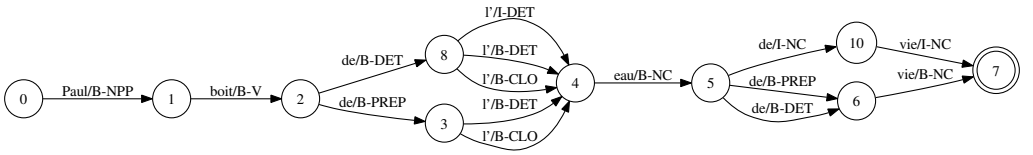


FIGURE 1 – Analyse lexicale dans le schéma BIO

3 Quelques expériences

Nous avons réalisé un certain nombre d'expériences sur le français, dans lesquelles notre corpus de référence était le corpus arboré de Paris 7 (Abeillé *et al.*, 2003), dans sa version de juin 2010. Ce corpus est composé d'articles journalistiques du Monde. Les mots composés sont marqués et considérés comme des unités lexicales. Nous avons utilisé les CRF (Lafferty *et al.*, 2001) et le logiciel *Wapiti* (Lavergne *et al.*, 2010) pour leur apprentissage et leur application. Les ressources lexicales et grammaticales étaient appliquées au moyen du logiciel *Unitex* [2].

Pour la segmentation lexicale, les ressources linguistiques externes sont composées de dictionnaires électroniques de mots simples et de mots composés, ainsi que de grammaires locales d'unités polylexicales semi-figées, compilées en transducteurs finis pour leur application. Nous avons observé des gains significatifs à la fois en terme de reconnaissance des mots composés (environ +5 points de F-mesure, pour des valeurs absolues autour de 70-80%) et d'étiquetage grammatical (+0.5 point, pour des valeurs d'absolues autour de 94%), par rapport au même segmenteur-étiqueteur sans ressources linguistiques externes.

Pour la segmentation syntaxique, nos ressources comprenaient, en plus, des grammaires locales de constituants syntaxiques simples. Nous avons montré, en particulier, leur grand intérêt sur des textes d'un autre domaine que le corpus de référence. Dans notre cas, nous avons travaillé sur un texte littéraire et des rapports parlementaires. Nous avons observé des gains de l'ordre de 8 points en terme de reconnaissance des constituants syntaxiques.

4 Conclusions

Ces expériences, comme celles de (Tellier et Dupont, 2013), montrent qu'il est avantageux de prendre en compte les informations véhiculées par un dispositif symbolique comme les automates à états finis dans des systèmes d'apprentissage automatique statistique comme les CRF. Les deux approches sont en effet complémentaires : les automates capturent des propriétés globales de la séquence, qui sont lexicalisées dans le codage BIO et injectées dans le CRF, qui à son tour évalue la confiance qu'il peut leur accorder pour le choix de l'étiquette finale.

2. http ://igm.univ-mlv.fr/˜unitex/.

© ATALA

Références

ABEILLÉ, A., CLÉMENT, L. et TOUSSENEL, F. (2003). Building a treebank for french. *In* ABEILLÉ, A., éditeur : *Treebanks*. Kluwer, Dordrecht.

BLANC, O., CONSTANT, M. et WATRIN, P. (2007). Segmentation in super-chunks with a finite-state approach. *In Proceedings of the Workshop on Finite-State Methods for Natural Language Processing (FSMNLP'07)*.

CONSTANT, M. et TELLIER, I. (2012). Evaluating the impact of external lexical resources into a crf-based multiword segmenter and part-of-speech tagger. *In Proceedings of the 8th conference on Language Resources and Evaluation (LREC'12)*.

DENIS, P. et SAGOT, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. *In Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC'09)*.

LAFFERTY, J., MCCALLUM, A. et PEREIRA, F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *In Proceedings of the Eighteenth International Conference on Machine Learning (ICML'01)*, pages 282–289.

LAVERGNE, T., CAPPÉ, O. et YVON, F. (2010). Practical very large scale CRFs. *In Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, pages 504–513.

RAMSHAW, L. A. et MARCUS, M. P. (1995). Text chunking using transformation-based learning. *In Proceedings of the 3rd Workshop on Very Large Corpora*, pages 88 – 94.

TELLIER, I. et DUPONT, Y. (2013). Apprentissage symbolique et statistique pour le chunking : comparaison et combinaisons. *In actes de TALN'2013*.

© ATALA

Determining order restrictions in recursive modification using compositionality in distributional semantics

Eva Maria Vecchi Roberto Zamparelli Marco Baroni
Centro Interdipartimentale Mente/Cervello, Universitá di Trento
evamaria.vecchi@unitn.it, roberto.zamparelli@unitn.it, marco.baroni@unitn.it

RÉSUMÉ

Compositionnalité en sémantique distributionnelle et ordonnancement des adjectifs

Dans cette présentation, nous proposons une approche non supervisée utilisant des méthodes compositionnelles en sémantique distributionnelle pour déterminer les restrictions qui s'appliquent à l'ordre des adjectifs d'un même syntagme nominal. Plus précisément, nous cherchons à distinguer automatiquement les syntagmes dont l'ordre des adjectifs est souple de ceux dont les adjectifs suivent un ordre fixe. Nous explorons plusieurs mesures qualitatives et quantitatives extraites de la représentation distributionnelle d'un syntagme qui sont toutes susceptibles d'indiquer une restriction de l'ordre des mots. Nous arrivons effectivement à distinguer les deux cas en utilisant principalement le degré de modification des adjectifs. Actuellement, notre approche ne concerne que l'anglais, mais, à notre avis, la méthodologie et les résultats peuvent s'adapter au français. Ce travail offre de nouvelles perspectives sur les propriétés sémantiques qui déterminent l'ordre des adjectifs, comblant ainsi l'écart entre syntaxe et sémantique distributionnelle.

ABSTRACT

In this talk, we propose an unsupervised approach to determine restrictions in adjective ordering through the use of compositional methods in distributional semantics. Specifically, we focus on being able to automatically distinguish phrases in which there is flexibility in the adjective ordering from those bound to a rigid order. We explore a number of measures extracted from the distributional representation of a phrase, both qualitative and quantitative, which may indicate a word order restriction. We find that we are able to distinguish the two cases of ordering significantly based primarily on the degree of modification of the adjectives. Currently, this approach has only been implemented in English, yet we predict that the methodology and results pertain naturally to French. This method offers fresh insight into the semantic properties that determine adjective ordering, building a bridge between syntax and distributional semantics.

MOTS-CLÉS : Sémantique distribuionnelle, Compositionnalité, Ordre des mots.

KEYWORDS: Distributional Semantics, Compositionality, Word-order.

1 Introduction

A prominent approach for representing the meaning of a word in Natural Language Processing (NLP) is to treat it as a numerical vector that codes the pattern of co-occurrence of that word with other expressions in a large corpus of language (Sahlgren, 2006; Turney et Pantel, 2010). This approach to semantics (sometimes called *distributional semantics*, DS) naturally captures

 © ATALA

collocations, scales well to large lexicons and does not require words to be manually disambiguated (Schütze, 1997). Until recently, however, this method had been almost exclusively limited to the level of content words (nouns, adjectives, verbs), and had not directly addressed the problem of *compositionality* (Frege, 1892; Montague, 1970; Partee, 2004), the crucial property of natural language which allows speakers to derive the meaning of a complex linguistic constituent from the meaning of its immediate syntactic subconstituents.

Several recent proposals have strived to extend distributional semantics with a component that also generates vectors for complex linguistic constituents, using compositional operations in the vector space (Baroni et Zamparelli, 2010; Guevara, 2010; Mitchell et Lapata, 2010; Grefenstette et Sadrzadeh, 2011; Socher *et al.*, 2012). All of these approaches manage to construct distributional representations for novel phrases, starting from the corpus-derived vectors for their lexical constituents and exploiting the geometric quality of the representation. Recent studies have shown that such methods are able to capture complex semantic information of adjective-noun (AN) phrases, such as distinguishing types of modification (Boleda *et al.*, 2012) and even detecting semantic deviance in novel phrases (Vecchi *et al.*, 2011, 2013). Further, such methods are naturally recursive. This aspect is appealing for modeling recursive modification since extracting meaningful representations for recursive phrases from a corpus will likely result in a problem of sparsity as most of them are hardly seen in a corpus.

However, the presence of *modifier ordering restrictions* has often been discussed in the theoretical linguistic literature (Sproat et Shih, 1990), and has become one of the key ingredients of the 'chartographic' approach to syntax. In this paradigm, the ordering is derived by assigning semantically different classes of modifiers to the specifiers of distinct functional projections, whose sequence is hard-wired. Various types of movement can then lead to a principled rearrangement of the linear order of the modifiers (Cinque, 2002, 2010; Steddy et Samek-Lodovici, 2011). Given that the existence of rigid ordering of adjectives is attributed to the semantic classes of modifiers, a good semantic representation should be able to capture restrictions in ordering due to their semantics.

We propose an unsupervised approach to determine restrictions in adjective ordering through the use of compositional methods in DS. Specifically, we focus on being able to automatically distinguish cases in which there is flexibility in the adjective ordering (e.g. *total estimated population* vs. *estimated total population*) from those bound to a rigid order (e.g. *rapid social change* vs. **social rapid change*). Being able to detect ordering restrictions proposes a number of benefits to various fields. In NLP, for example, information about such restrictions can help determine the plausibility of possible parses of recursive phrases. Gaining a better understanding of the semantic restrictions may provide insight to the cognitive processing of phrases. Finally, it stimulates the possibility to learn how distributional semantics and syntax can overlap to provide a complete representation of natural language.

2 Approach

We begin with a semantic space in which our vocabulary (adjectives, nouns and ANs) is represented as vectors whose dimensions are a pattern of co-occurrences with content words (nouns, verbs, adjectives and adverbs) extracted from a large corpus. We then generate a set of adjective-adjective-nouns (AANs) composed of the most frequent adjectives and nouns. These

 © ATALA

phrases are divided into two types of adjective ordering : Flexible ordering (FO) items occur frequently as both A_xA_yN and A_yA_xN, and Rigid ordering (RO) items occur frequently in only one order, while the other is unattested. These corpus-observed AANs are then mapped onto the semantic space.

For the study introduced here, we focus on two composition functions [1] with high performance in a number of semantic tasks : the weighted additive function (*add*) from (Mitchell et Lapata, 2010) and the lexical function model (*lfm*) from (Baroni et Zamparelli, 2010) [2]. We apply the functions recursively and generate predicted vectors for the AANs.

Our first analysis is a qualitative inspection of the nearest neighbors of the AAN vectors, both corpus-observed and predicted, as represented in the semantic space. To our knowledge, this is the first study to use distributional representations of recursive modification ; therefore we must first determine if the vector representations are semantically coherent objects. We then explore whether we can detect a difference in adjective ordering based on the distance relationship between the AAN vectors and each of its subparts. Specifically, we consider the distance between the AAN vector and (i) the component N vector, (ii) the component A vectors, and (iii) the component AN vectors. We predict that in cases of flexible ordering, the two adjectives will have a similarly strong effect on the noun, thus transforming the meaning of the noun equivalently in the direction of both adjectives. On the other hand, in cases of rigid ordering, we predict that one adjective will dominate the meaning of the phrase, thus distorting the meaning of the noun significantly in the direction of one adjective.

3 Results

First, in a qualitative analysis of the nearest neighbors of the AANs, we find that corpus-extracted AAN vectors appear to be meaningful, semantically coherent objects. We also find that the neighbors for the AANs represent an interesting variety of types of semantic similarity. For example, the nearest neighbors to the observed vector for *medieval old town* are the phrases *fascinating town*, *impressive cathedral* and *medieval street*. We also find that both the *add* and *lfm* functions approximate the observed vectors quite well, obtaining a mean (cosine) distance from the observed equivalent over 0.5 for both functions.

The analysis of the distance relationship between the AANs and their subparts provided a clear, significant distinction between FO and RO phrases. Given an A_xA_yN, FO AANs are closer to their A_x, A_xN and the component N than RO AANs. Further, RO AANs remain significantly closer to their A_y. These results imply that the strength of modification of the adjective on the noun is a strong indicator of the type of ordering present. RO AANs are heavily modified by A_y, distorting the meaning of the head noun drastically, and only maintaining a slight modification from A_x. This supports the hypothesis that the A_yN is in fact lexicalized to some degree. FO AANs, on the other hand, maintain the constitution of the head noun while being modified only slightly by both adjectives, almost equivalently.

Further exploration into being able to predict word order given an adjective, adjective and noun remains the next step in this investigation.

1. Since the multiplicative function (Mitchell et Lapata, 2010) is insensitive to word order, we do not test it here.
2. For space reasons, we do not provide equations for these functions. See the referenced literature for details.

 © ATALA

References

BARONI, M. et ZAMPARELLI, R. (2010). Nouns are vectors, adjectives are matrices : Representing adjective-noun constructions in semantic space. *In Proceedings of EMNLP*, pages 1183–1193, Boston, MA.

BOLEDA, G., VECCHI, E., CORNUDELLA, M. et MCNALLY, L. (2012). First-order vs. higher-order modification in distributional semantics. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

CINQUE, G., éditeur (2002). *Functional Structure in DP and IP - The Carthography of Syntactic Structures*, volume 1. Oxford University Press.

CINQUE, G. (2010). *The syntax of adjectives : a comparative study*. MIT Press.

FREGE, G. (1892). Über sinn und bedeutung. *Zeitschrift fuer Philosophie un philosophische Kritik*, 100.

GREFENSTETTE, E. et SADRZADEH, M. (2011). Experimental support for a categorical compositional distributional model of meaning. *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, UK.

GUEVARA, E. (2010). A regression model of adjective-noun compositionality in distributional semantics. *In Proceedings of the ACL GEMS Workshop*, pages 33–37, Uppsala, Sweden.

MITCHELL, J. et LAPATA, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

MONTAGUE, R. (1970). Universal grammar. *Theoria*, 36(3):373–398.

PARTEE, B. (2004). Compositionality. *In Compositionality in Formal Semantics : Selected Papers by Barbara H. Partee*. Blackwell, Oxford.

SAHLGREN, M. (2006). *The Word-Space Model*. Dissertation, Stockholm University.

SCHÜTZE, H. (1997). *Ambiguity Resolution in Natural Language Learning*. CSLI, Stanford, CA.

SOCHER, R., HUVAL, B., MANNING, C. D. et NG, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. *In Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

SPROAT, R. et SHIH, C. (1990). The cross-linguistics distribution of adjective ordering restrictions. *In* GEORGOPOULOS, C. et R., I., éditeurs : *Interdisciplinary approaches to language : essays in honor of Yuki Kuroda*, pages 565–593. Kluver, Dordrecht.

STEDDY, S. et SAMEK-LODOVICI, V. (2011). On the ungrammaticality of remnant movement in the derivation of greenberg's universal 20. *Linguistic Inquiry*, 42(3):445–469.

TURNEY, P. et PANTEL, P. (2010). From frequency to meaning : Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

VECCHI, E. M., BARONI, M. et ZAMPARELLI, R. (2011). (Linear) maps of the impossible : Capturing semantic anomalies in distributional space. *In Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 1–9. Association for Computational Linguistics.

VECCHI, E. M., MARELLI, M., ZAMPARELLI, R. et BARONI, M. (2013). Spicy adjectives and nominal donkeys : Capturing semantic deviance using compositionality in distributional spaces. Submitted.

 © ATALA

Inférence de règles déductives par abduction

Mathieu Lafourcade, Manel Zarrouk, Alain Joubert
LIRMM, 161, rue ADA 34392 Montpellier Cedex 5
mathieu.lafourcade@lirmm.fr, manel.zarrouk@lirmm.fr, alain.joubert@lirmm.fr

RÉSUMÉ

Les réseaux lexico-sémantiques sont un type de ressources majeur en TAL. Indépendamment des stratégies de construction utilisées, inférer automatiquement des règles avec lesquelles de nouvelles relations peuvent être produites est une approche possible pour améliorer la couverture et la qualité globale de la ressource. Ce type de règle peut également être utilisé avec profit pour l'analyse sémantique de textes. Dans ce contexte, un moteur d'inférences a pour tâche d'explorer le contenu du réseau lexical et d'en extraire des règles potentielles. Seules les règles validées par des contributeurs ou des experts sont mémorisées et appliquées afin d'enrichir le réseau et en retour de découvrir de nouvelles règles. La nature polysémique des termes de la ressource et les nombreux silences sont autant d'obstacle à la validation automatique des règles candidates, cependant des évaluations statistiques combinées à des traits symboliques permettent d'obtenir et de filtrer les règles douteuses efficacement.

ABSTRACT

Deductive rule inference through abduction

In NLP, lexical semantic networks represent a major and crucial type of resources. Nonetheless the construction strategy used, automatically inferring rules with which new relations can be produced is a possible way to improve coverage and global quality of the resource. Such rules can be used likewise profitably for semantic analysis of texts. Within that context, an inference engine has for goal to explore the contents of the lexical network and extract potential rules. Only rules validated by contributors or experts are memorized and applied latterly as patterns to enrich the network and so on discover new rules. The polysemy of terms in the resource and the presence of silence represent obstacles to the automatic validation of the candidate rules. However, statistical evaluations combined to some symbolic features, allow efficiently to have and to filter out the dubious rules.

MOTS-CLÉS: inférence de règles, enrichissement, réseau lexico-sémantique, traits symboliques, filtrage statistique.
KEYWORDS: rule inference, enrichment, semantic lexical network, symbolic features, statistical filtering.

© ATALA

1 Contexte et motivation

Le projet JeuxDeMots (Lafourcade, 2007) vise à produire un réseau lexico-sémantique à l'aide d'approches contributives dont notamment des jeux en ligne. L'inférence d'associations dans le réseau est un moyen efficace de le consolider, à savoir ajouter de nouvelles relations et de détecter celles possiblement erronées. Dans (Zarrouk et al, 2013) un schéma d'inférence déductif a été proposé afin de produire des associations dans un réseau lexical. Les règles d'inférence, qui sont des schémas avec une seule variable, peuvent être utilisées pour l'amélioration de la ressource, mais également avec profit lors d'une analyse sémantique de texte.

La plupart des ressources existantes de type réseau lexico-sémantique ont été construites manuellement. Des outils sont souvent utilisés pour la vérification de la cohérence, mais la tâche reste coûteuse en temps et en prix. Les approches automatisées à partir de corps sont généralement limitées à la cooccurrence des termes, l'extraction de relations sémantiques précises restant difficile. L'externalisation ouverte (crowdsourcing) émerge en particulier avec l'avènement de Amazon Mechanical Turk et plus largement avec Wikipédia. Wordnet ((Miller, 1990) et (Fellbaum, 1998)) est un réseau lexical de l'anglais fondé sur des pouvant être considérés comme des concepts. EuroWordnet (Vossen, 1998), une version multi-langues de Wordnet et WOLF (Sagot, 2008), une version française de Wordnet, sont issus de croisements automatiques de Wordnet avec d'autres ressources lexicales suivi d'une vérification manuelle partielle. Dans le domaine de l'intelligence artificielle, Cyc (Lenat, 1995) est une base de connaissances très riche ayant demandé un effort manuel particulièrement important. BabelNet (Navigli, 2012) est un réseau lexical multilingue construit automatiquement à partir des cooccurence de termes de Wikipédia. HowNet (Dong, 2008) est un autre exemple d'une base de connaissances bilingue (anglais et chinois) contenant des relations sémantiques entre des formes de mots, des concepts et des attributs. (Harabagui, 1998) a réalisé des inférences à partir de textes en se basant sur WordNet comme base de connaissance. L'évaluation de règles inférées depuis des corpus reste difficile comme le souligne Zeichner (2012). A notre connaissance, il n'existe pas de travaux d'inférence endogène de règles de déduction à partir de réseaux lexicaux construits par externalisation ouverte.

Un réseau lexico-sémantique peut contenir des formes de mots, des termes, des concepts, des expressions composées et idéalement des sens de termes (des usages). Les usages ne sont pas pour autant nécessairement totalement désambigüisés et peuvent avoir eux-mêmes plusieurs raffinements. Par exemple, *frégate* peut être un *oiseau* ou un *navire*. Une *frégate>bateau* peut être distinguée comme un *navire moderne* ou un *navire à voiles ancien*. Les concepts présents dans le réseau peuvent prendre la forme de traits sémantiques généraux, par une relation nommée SEM non ambigus sous la forme de symboles (VIVANT, HUMAIN, ARTÉFACT, OBJET NATUREL, LIEU, TEMPS, etc.). Dans le contexte d'une approche collaborative, un réseau lexical doit être considéré comme en cours de construction et l'est en pratique. Certains usages ainsi que de nombreuses relations peuvent manquer et il est difficile de connaître l'état de silence de chaque entrée.

Il s'agit alors de produire automatiquement des règles d'inférence déductives par inspection du contenu du réseau lexical. Une règle, une fois validée, sera appliquée afin de produire de nouvelles relations dans le réseau. La polysémie du réseau ainsi que les silences sont les

 © ATALA

principaux obstacles à l'identification statistique de règles correctes. Par contre, l'information symbolique quand elle est présente constitue une aide précieuse au filtrage des règles candidates.

2 Expérience et résultats

A partir du réseau lexical JeuxDeMots nous avons cherché à produire des règles inductives de la forme « *Si x est une sorte de B alors x a une relation R avec C* » (que l'on notera : $\forall x, is\text{-}a\ x, B \Rightarrow R(x, C)$). L'interprétation d'une telle règle est qu'elle doit être vraie pour la plupart des *x*, modulo les exceptions qui devront rester en nombre limité pour être considérées comme telles. Il s'agit depuis des informations potentielles présentes dans le réseau de déterminer celles qui sont systématiquement valides. L'approche abductive a consisté à énumérer systématiquement les propriétés existantes pour chaque terme ayant au moins un hyperonyme. Il s'agit donc de trouver dans le réseau lexical des exemples fournissant un support à l'abduction de règles. Pour chaque terme B, on effectue les comptages suivants :

- #E : le nombre de *x* (tel que is-a(x, B) > 0), il s'agit des termes concernés par la règle ;
- #ER$^+$: le nombre de *x* (is-a B) renseignés pour R et ayant R(x,C) > 0 (les exemples positifs) ;
- #ER$^-$: le nombre de *x* (is-a B) renseignés pour A et ayant R(x, C) < 0 (les contre-exemples).

On ne considère que les règles candidates où le nombre d'exemples #ER$^+$ dépasse un seuil donné (>= 1). L'approche produit un grand nombre de règles candidates, à savoir plus de 330 000, pour 91 896 relations x is-a B présentes dans le réseau.

# exemples min (#min)	1	3	5	**10**	20	40
# candidats	336 000	3 593	1 718	**649**	330	92

La quantité de silence ($\#S = \#E - (\#ER^+ + \#ER^-)$) pour une règle est le nombre d'exemples relevant de la règle candidate qui ne sont pas du tout renseignés pour la relation R. Nous en déduisons le taux de véracité TV ($\#S\ \ \#ER^+$) de la règle candidate et un taux de fausseté TF ($\#S\ \ ER^-$). Par exemple, pour la règle candidate *Tout oiseau vole* $\forall x, is\text{-}a\ x, oiseau \Rightarrow agent(x, voler(déplacement\ aérien))$, nous avons TV = 87% et TF=0.7%. C'est à dire que pour les termes ayant pour hyperonyme *oiseau*, ceux renseignés pour la relation *agent* sont très majoritairement reliés à *voler* et qu'un petit nombre est négativement relié à *voler* (*autruche, émeu, casoar, pingouin*). Empiriquement, il a été constaté que la tolérance de 1% d'exceptions combinée à au maximum 50% de silence fournissait le F1-score optimal de 80% (maximum de règles fausses et minimum de règles correctes éliminées) pour un nombre minimal d'exemples #min = 10.

La polysémie est un facteur de règles erronées dont l'effet est grandement atténué par la prise en considération de traits symboliques présents dans le réseau. Par exemple, nous avons obtenu la règle candidate fausse $\forall x, is\text{-}a\ x, métier \Rightarrow partie(x, cou)$ car la plupart des termes de métier fusionnent à la fois l'activité et la personne. En considérant les traits sémantiques majoritaires dans le réseau lexical, il est possible d'abduire la règle $\forall x, is\text{-}a\ x, métier \wedge sem(x, \text{HUMAIN}) \Rightarrow partie(x, cou)$, qui s'interprète de la façon

© ATALA

suivante *«si A est un terme de métier et a le trait sémantique* HUMAIN *alors A peut avoir comme partie* COU *»*. Pour le terme *métier à risque*, le trait sémantique HUMAIN ne sera pas trouvé et la règle ne sera pas appliquée.

La prise en compte de symboles sémantiques associés aux termes dans le réseau lexical de JeuxDeMots permet d'obtenir un F1-score de 93%. Les cas d'échec sont les cas rares non renseignées aboutissant à des généralisations abusives, et qui concernent un nombre de règles réduit (8%) où il est possible de solliciter les contributeurs afin d'en énumérer les exceptions. Au final, l'abduction consiste ici à débusquer et transformer des informations indiquant une possibilité *(des oiseaux peuvent voler)* en règles généralement vraies *(tout oiseau peut voler).*

Références

BLANCO, E. et MOLDOVAN, D. (2011). A model for composing semantic relations. *Ninth International Conference on Computational Semantics (IWCS'11),* Oxford, United Kingdom, pages 45–54.

DONG, Z. et DONG, Q. (2006). *HowNet and the Computation of Meaning.* WorldScientific, London.

FELLBAUM, C. et MILLER, G. (1998). *(eds) WordNet.* The MIT Press.

HARABAGIU, S. et MOLDOVAN, D. (1998). Knowledge processing on an extended wordnet. *WordNet : An Electronic Lexical Database, MIT Press.*, pages 381–405.

LAFOURCADE, M. (2007). Making people play for lexical acquisition. *In Proc. SNLP 2007, 7th Symposium on Natural Language Processing.* Pattaya, Thailande, 13-15 December 2007, page 8 p.

LENAT, D. (1995). Cyc : A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

LIEBERMAN, H., SMITH, D. A. et TEETERS, A. (2007). Common consensus : a web-based game for collecting commonsense goals. *In Proc. of IUI*, Hawaii., page 12 p.

MARCHETTI, A., TESCONI, M., RONZANO, F., MOSELLA, M. et MINUTOLI, S. (2007). Semkey : A semantic collaborative tagging system. *in Procs of WWW2007,* Banff, Canada, page 9 p.

MIHALCEA, R. et CHKLOVSKI, T. (2003). Open mindword expert : Creating large annotated data collections with web users help. *In Proceedings of the EACL 2003, Workshop on Linguistically Annotated Corpora (LINC 2003)*, page 10 p.

MILLER, G., BECKWITH, R., FELLBAUM, C., GROSS, D. et MILLER, K. (1990). Introduction to wordnet : an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

NAVIGLI, R. et PONZETTO, S. (2012). Babelnet : Building a very large multilingual semantic network. *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* Uppsala, Sweden, 11-16 July 2010, pages 216–225.

SAGOT, B. et FIER, D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues. *TALN 2008*, Avignon, France, 2008., page 12.

 © ATALA

von AHN, L. et DABBISH, L. (2008). Designing games with a purpose. *Communications of the ACM*, 51(8):58–67.

VOSSEN, P. (1998). *Eurowordnet: a multilingual database with lexical semantic networks.* Kluwer Academic Publishers, Norwell, MA, USA, page 200.

ZARROUK, M., LAFOURCADE, M and JOUBERT, A. (2013) Inference and reconciliation in a lexical-semantic network *In proc of 14th International Conference on Intelligent Text Processing and Computational Linguistic (CILING-2013)*, University of the Aegean, Samos, Greece, March 24–30, 2013, 13 p.

ZEICHNER, N., BERANT, J. and DAGAN, I. (2012) Crowdsourcing Inference-Rule Evaluation *ACL 2012*, 5 p.

ZESCH, T. et GUREVYCH, I. (2009). Wisdom of crowds versus wisdom of linguists measuring the semantic relatedness of words. *Natural Language Engineering, Cambridge University Press.*, pages 25–59.

© ATALA

Probabiliser les Grammaires de Propriétés

Philippe Blache & Stéphane Rauzy
Aix-Marseille Université, CNRS, LPL
5 Avenue Pasteur, 13100 Aix-en-Provence
{blache;rauzy}@lpl-aix.fr

RÉSUMÉ
Nous présentons dans cet article une description des principaux facteurs de complexité de l'analyse syntaxique en *Grammaire de Propriétés* et proposons une stratégie permettant d'en réduire le coût à l'aide d'une stratégie hybride intégrant un contrôle symbolique à un processus probabiliste.

ABSTRACT
Probabilistic Property Grammars

We describe in this paper the main complexity factors of parsing in *Property Grammars*. A strategy in presented, making it possible to reduce the costs thanks to the integration of a symbolic control into a probabilistic process.

MOTS-CLÉS : Grammaires de Propriétés, satisfaction de contraintes, treebank.

KEYWORDS: Property Grammars, constraint satisfaction, treebanks.

Les *Grammaires de Propriétés* (Blache, 2001) font partie de ces formalismes visant à l'implantation de théories linguistiques généralistes Elles permettent d'une part la représentation d'informations provenant de domaines différents (prosodie, discours, syntaxe, etc.) et offrent d'autre part la possibilité de traiter des énoncés non canoniques (typiquement, la langue parlée). L'intérêt de ces approches réside dans la capacité de décrire finement des constructions ou des phénomènes complexes, permettant de prendre en compte le langage dans sa globalité.

1 L'analyse en Grammaires de Propriétés : un processus complexe

Si les *Grammaires de Propriétés* (dorénavant GP) offrent un avantage en termes d'expressivité, leur traitement automatique reste cependant un processus complexe. Nous nous limiterons dans ce qui suit au cas de l'analyse syntaxique. Dans ce domaine, l'analyse consiste à associer à chaque catégorie l'ensemble des propriétés qui la caractérise, ces propriétés étant des relations de divers types entre les constituants : ordre linéaire ($\prec$), obligation de cooccurrence ($\Rightarrow$), interdiction de cooccurrence($\not\Leftrightarrow$), dépendance ($\rightsquigarrow$), interdiction de répétition(*Unic*) et tête (*Oblig*).

Le mécanisme d'analyse consiste à identifier un ensemble de catégories (les constituants d'une unité) et d'en évaluer les propriétés. L'ensemble de constituants est appelé *affectation*, il est toujours possible d'en calculer les propriétés grâce au relâchement de contraintes : une propriété

 © ATALA

pourra être satisfaite ou enfreinte (permettant ainsi la description d'énoncés non grammaticaux).

L'analyse syntaxique en *GP* repose sur un processus en plusieurs étapes :

1. Identification des *affectations*
2. Evaluation des propriétés décrivant les relations entre objets dans cet espace
3. Inférence de nouvelles catégories

Au départ sont disponibles toutes les catégories lexicales de la phrase à analyser. Une affectation est un sous-ensemble de ces catégories : l'ensemble des affectations est donc l'ensemble des sous-ensembles de catégories. Dans l'usage (cette restriction étant adaptée à des langues comme le français), nous nous limitons à l'ensemble des catégories *adjacentes*.

	La	*principale*	*activité*	*du*	*domaine*	...
Exemple :	`Det`	`Adj`	`N`	`Prep`	`N`	
	`Pro`	`N`				
	`N`					

Les affectations sont ici `{Det}`; `{Det, Adj}`; `{Det, Adj, N}`; `{Pro, Adj}`, `{N, Prep, N}`, etc.

L'étape de calcul des propriétés consiste à évaluer pour chacune des affectations les propriétés mettant en relation les objets la concernant (Balfourier *et al.*, 2002). Par exemple, les propriétés de *linéarité*, de *dépendance* et d'*exigence* seront satisfaites pour l'espace `{Det, Adj}`.

L'étape suivante consiste à évaluer lesquelles des affectations correspondent à la description d'une catégorie syntagmatique. Il s'agit pour cela d'analyser l'ensemble des propriétés évaluées pour cette affectation (ensemble appelé également *caractérisation*). Chaque propriété étant propre à une catégorie syntaxique donnée, analyser la caractérisation consiste à rechercher si les propriétés qui la composent sont associées à une même catégorie syntagmatique. Si c'est le cas, cette catégorie est inférée et ajoutée à l'ensemble des catégories à partir duquel de nouvelles affectations seront créées. Le processus en trois étapes recommence tant que de nouvelles catégories peuvent être inférées.

La dernière étape concerne l'identification d'une solution parmi les ensembles d'observation. Une solution est un sous-graphe :

– qui couvre la totalité de l'input
– dans lequel chaque mot (et plus généralement chaque espace d'observation) correspond à une étiquette unique
– qui satisfait un nombre suffisant de propriétés (calcul prenant également en compte le poids des propriétés (Blache, 2012).

La stratégie d'analyse est donc particulièrement coûteuse et seules des heuristiques permettent d'en contrôler la mise en œuvre (limitation des espaces d'observation, utilisation de coins gauches, introduction d'un seuil de satisfaisabilité, etc.).

© ATALA

Propriétés	Réalisations				
	{Det,N}	{Det,N,AP}	{Det,N,PP}	{Det,N,AP,PP}	{N,NP}
[1] Det $\prec$ N	x	x	x	x	
[2] Det $\prec$ AP		x		x	
[3] Det $\prec$ PP			x	x	
[4] Det $\prec$ NP					x
[5] N $\prec$ PP			x	x	
[6] Oblig(N)	x	x	x	x	x
[7] Det $\rightsquigarrow$ N	x	x	x	x	
[8] PP $\rightsquigarrow$ N			x	x	
[9] AP $\rightsquigarrow$ N		x		x	
[10] Det $\Rightarrow$ N	x	x	x	x	x
[11] PP $\Rightarrow$ N			x	x	
[12] AP $\Rightarrow$ N		x		x	
[13] Unic(Det)	x	x	x	x	
[14] Unic(N)	x	x	x	x	
[15] N $\not\Leftrightarrow$ Pro	x	x	x	x	x
[16] N $\not\Leftrightarrow$ Npro	x	x	x	x	x
[17] AP $\not\Leftrightarrow$ Pro		x		x	
[18] AP $\not\Leftrightarrow$ Ssub		x		x	
[19] AP $\not\Leftrightarrow$ VPinf		x		x	
[20] PP $\not\Leftrightarrow$ Ssub			x	x	

2 Comment et quoi probabiliser ?

L'enrichissement de treebanks syntagmatiques du type FTB (Abeillé *et al.*, 2003) par des analyses en *GP* est un processus assez simple (Blache et Rauzy, 2012). Nous disposons pour chaque catégorie à la fois de toutes ses affectations effectivement réalisées ainsi que de leurs caractérisations. Parmi les informations pouvant être utiles au contrôle du processus d'analyse, la connaissance des propriétés satisfaites conjointement est importante. Le tableau suivant présente les caractérisations de 5 affectations du *NP* : les croix indiquent que la contrainte de la ligne courante est satisfaite pour l'affectation indiquée en colonne, une caractérisation est donc une colonne de croix.

Il est ainsi possible de déterminer des ensembles de contraintes satisfaites simultanément et d'en quantifier la cooccurrence par des probabilités. Etant donné un ensemble $\mathscr{P}$ de propriétés satisfaites, on peut déduire l'ensemble $\mathscr{P}'$ de propriétés qui seront le plus probablement satisfaites en même temps que $\mathscr{P}$. Remarquons qu'il est possible d'intégrer la violation de contraintes dans le modèle, au même niveau que la satisfaction : une contrainte peut en effet être systématiquement enfreinte dans un contexte particulier.

Dans l'exemple ci-dessus, la propriété *Det* $\prec$ *N* est systématiquement satisfaite en même temps que les propriétés *Oblig(N), Det* $\rightsquigarrow$ *N, Det* $\Rightarrow$ *N* , *Unic(Det), Unic(N), N* $\not\Leftrightarrow$ *Pro* et *N* $\not\Leftrightarrow$ *Npro*. Par ailleurs, cette même propriété est satisfaite pour le sous-ensemble de réalisations `{Det,N}`; `{Det,N,AP}`; `{Det,N,PP}`; `{Det,N,AP,PP}`

Il est donc possible, connaissant une propriété, d'une part de réduire la génération des affectations aux seuls ensembles effectivement réalisés et d'autre part d'identifier un sous-ensemble de propriétés dont la probabilité de satisfaction est élevée. En d'autres termes, la connaissance d'une propriété fournit des probabilités de réalisation d'un ensemble de catégories en même temps qu'une description syntaxique partielle. Ces informations permettent :

 © ATALA

– de limiter la génération des affectations aux seules affectations pertinentes (en d'autres termes prédire les catégories suivantes)
– d'initialiser une première caractérisation
– d'identifier les projections possibles (en d'autres termes, prédire la fermeture d'un syntagme)

Chacune de ces informations permet donc de réduire l'espace de recherche de chacune des étapes du processus d'analyse. Concrètement, la stratégie d'analyse se présente de la façon suivante :

$\{c_1 \dots c_n\}$
$\downarrow$ Recherche des propriétés $\mathscr{P}$ les plus probables pour l'affectation $\{c_1 \dots c_n\}$
$\mathscr{P}$
$\downarrow$ Recherche des propriétés $\mathscr{P}'$ cooccurentes de $\mathscr{P}$
$\mathscr{P}'$
$\downarrow$ Recherche de l'ensemble des affectations $\mathscr{A}$ associées à $\mathscr{P} \cup \mathscr{P}'$
$\mathscr{A}$
$\downarrow$ Recherche dans $\mathscr{A}$ de la catégorie (la transition) la plus probable
$\{c_1 \dots c_{n+i}\}$

Ce processus permet donc de contrôler la prédiction d'une catégorie plus précisément qu'en prenant en compte simplement la suite de catégories du contexte gauche, comme c'est le cas dans les stratégies probabilistes basées sur les grammaires syntagmatiques. De plus, ce processus permet d'analyser les inputs mal formés : une affectation, même si elle contient des propriétés non satisfaites, permettra de déterminer une probabilité de transition grâce à la satisfaction d'autres propriétés.

Références

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for french. *In* Abeillé, A., éditeur : *Treebanks*, Kluwer, Dordrecht.

Balfourier, J.-M., Blache, P. et Rullen, T. V. (2002). From Shallow to Deep Parsing Using Constraint Satisfaction. *In Proc. of the 6th Int'l Conference on Computational Linguistics (COLING 2002)*.

Blache, P. (2001). *Les Grammaires de Propriétés : Des contraintes pour le traitement automatique des langues naturelles*. Hermès.

Blache, P. (2012). Estimating constraint weights from treebanks. *In Proceedings of CSLP*.

Blache, P. et Rauzy, S. (2012). Hybridization and treebank enrichment with constraint-based representations. *In Workshop on Advanced Treebanking*.

© ATALA

Utiliser une base distributionnelle pour filtrer un dictionnaire de synonymes

François Morlane-Hondère
CLLE-ERSS, Université de Toulouse - Le Mirail, 5, allées Antonio Machado - Toulouse Cedex 9
francois.morlane@univ-tlse2.fr

RÉSUMÉ

Cette étude vise à mettre en lumière l'intérêt qu'il peut y avoir à se servir d'une ressource générée par analyse distributionnelle automatique pour orienter les résultats fournis par un dictionnaire de synonymes. En croisant une base distributionnelle calculée à partir d'un corpus constitué d'articles de l'encyclopédie Wikipédia et le Dictionnaire Électronique des Synonymes du CRISCO, nous montrons qu'une partie seulement des synonymes proposés pour un mot donné partagent ses contextes d'apparition. Nous mettons au jour plusieurs raisons qui expliquent ce phénomène. Nous montrons ensuite que ce décalage s'observe différemment selon la nature du corpus qui a permis de calculer la base distributionnelle qui sert à filtrer le dictionnaire. Cela signifie que la nature du corpus oriente le type de synonymes filtrés par la base distributionnelle. Nous envisageons d'appliquer ce principe dans un système de réorganisation dynamique des synonymes du dictionnaire du CRISCO.

ABSTRACT

Using distributional analysis for synonym selection

In this study, we aim to highlight the benefits of using a distributional resource to improve the sorting of the synonyms contained in a dictionary. We compare a distributional resource which was created from a Wikipedia corpus and the Electronic Dictionary of Synonyms of CRISCO. We show that only a part of a given word's synonyms share its contexts in the Wikipedia corpus. We find several explanations for this phenomenon. Then, we compare the dictionary with other distributional resources and we show that the nature of the corpus affects the range of the discrepancy. That means that the nature of the corpus has an influence on the kind of synonyms that appear in the distributional resources. We plan to use this principle for the automatic reorganization of the dictionary's synonyms.

MOTS-CLÉS : analyse distributionnelle, synonymie, corpus, sémantique lexicale.

KEYWORDS: distributional analysis, synonymy, corpus, lexical semantics.

1 Introduction

Cette étude s'inscrit dans la problématique globale de l'évaluation des bases lexicales générées par analyse distributionnelle automatique. Développé dans Harris (1968), le principe de l'analyse distributionnelle (désormais *AD*) s'appuie sur l'idée selon laquelle le degré de similarité sémantique entre deux mots peut se mesurer sur la base du recouvrement de leurs contextes syntaxiques. Il a été automatisé dès le début des années 90 et a été notamment utilisé par la

 © ATALA

suite pour assister la création d'ontologies (Grefenstette, 1994; Habert, 1998). Toutefois, les limitations propres aux outils d'analyse et le manque de données ne permettait réellement que d'entrevoir les possibilités de la méthode distributionnelle.

Couplé à l'arrivée d'analyseurs syntaxiques dits "robustes", l'accroissement de la quantité de textes accessibles au format électronique a été un facteur important dans la popularisation de la méthode distributionnelle. Ce phénomène a en effet permis de répondre à la nécessité pour l'AD de disposer de grandes quantités de données. Des ressources constituées d'archives de journaux, d'œuvres littéraires ou de textes issus du Web permettent aujourd'hui d'appliquer le principe distributionnel à grande échelle, sur des corpus de différentes natures.

Cette diversification rend encore plus prégnante la question du type de relations qui sont mises au jour ainsi que l'impact du corpus sur les bases distributionnelles générées. Différents travaux ont cherché à évaluer l'efficacité des méthodes distributionnelles pour le repérage des relations lexicales en employant des lexiques de référence (Turney, 2008; Baroni et Lenci, 2011) ou des tâches sémantiques (Baroni et Lenci, 2010). Dans des études précédentes, nous nous sommes notamment intéressés au cas de l'antonymie (Morlane-Hondère et Fabre, 2010) et de la méronymie (Morlane-Hondère et Fabre, 2012). Tous ces travaux confirment la grande diversité des relations que peut détecter l'AD, et montrent aussi la nécessité de mieux comprendre sous quelles conditions le critère distributionnel opère.

L'aspect évaluatif est destiné à prendre davantage d'importance avec la tendance actuelle à puiser des textes sur le Web (Turney, 2008; Baroni et Lenci, 2010) ou à fusionner plusieurs ressources de natures différentes (Baroni et Lenci, 2010). Cette démarche se fait au détriment d'une réflexion pourtant nécessaire autour de la caractérisation des corpus pour le calcul distributionnel. On sait en effet que les rapprochements produits par l'AD sont le reflet des fonctionnements des mots en corpus. Or, générer des bases distributionnelles à partir de corpus hétérogènes et/ou constitués de textes dont la nature est inconnue peut compliquer l'évaluation des ressources générées en gênant l'interprétation des résultats produits. Cela concourt à augmenter l'aspect *boîte noire* de la méthode, qui perd par conséquent de son intérêt du point de vue de l'analyse linguistique.

Nous proposons ici d'aborder cette double problématique de la caractérisation des résultats produits par cette méthode ainsi que celle de l'influence des corpus qu'elle prend en entrée à travers l'étude des effets du filtrage d'un dictionnaire de synonymes par une ressource distributionnelle. Notre démarche s'appuie sur l'hypothèse selon laquelle une base distributionnelle peut permettre de *sélectionner* les synonymes d'un dictionnaire en présentant à l'utilisateur ceux qui sont les plus immédiatement substituables à un mot donné (section 2). Après avoir présenté les ressources distributionnelles utilisées ainsi que le protocole que nous avons mis en œuvre – sections 3 et 4 –, nous procédons à une analyse linguistique des effets du filtrage par une base distributionnelle calculée à partir d'un corpus encyclopédique (section 5). Nous décrivons ensuite l'impact de la nature du corpus sur le filtrage des synonymes en croisant le dictionnaire et deux autres bases distributionnelles calculées à partir de corpus journalistique et littéraire (section 6).

2 Intérêts du filtrage

Le lexique de synonymes avec lequel nous travaillons est le Dictionnaire électronique des synonymes (*DES* ; Manguin *et al.*, 2004), développé au sein du CRISCO (Université de Caen). Il regroupe les synonymes relevés dans sept dictionnaires (dictionnaires analogiques et dictionnaires

 © ATALA

de synonymes) – le *Bailly*, le *Benac*, le *Du Chazaud*, le *Guizot*, le *Lafaye*, le *Larousse* et le *Robert*. La version dont nous disposons comporte 389 182 paires de noms (52 %), verbes (26 %) et adjectifs (22 %) [1], mots simples ou syntagmes. Il est à noter que les paires ont été symétrisées (pour une relation *A/B* a été générée une relation *B/A*) : le dictionnaire compte donc 194 576 relations de synonymie réciproques. Le DES étant régulièrement mis à jour, il comptait, au 25 octobre 2012, 200 849 relations réciproques. Ce dictionnaire a été précédemment employé dans des travaux comme ceux de Bourigault et Galy (2005) ou Morlane-Hondère et Fabre (2012) pour évaluer des bases distributionnelles.

Lancée en 1998, la version en ligne du DES a connu un succès grandissant, si bien qu'en octobre 2012, le site reçoit 200 000 requêtes par jour (Orosemane, 2012). Le dictionnaire peut être consulté soit *via* le site du CRISCO [2], soit sur la plate-forme CNRTL [3]. Or, ce lexique, consultable sur le Web, constitue pour de nombreux utilisateurs un outil d'aide à la rédaction. Manguin (2005) rapporte qu'il est notamment utilisé comme tel par les administrations québécoises et suisses, qui sont parmi les plus grands utilisateurs du DES. L'article de Orosemane (2012) témoigne également de son utilisation dans le milieu journalistique. La profusion de synonymes fournis par le DES pour un mot donné est toutefois perçue, par certains utilisateurs, comme une gêne à son utilisabilité (Thot Cursus, 2012).

L'ordre dans lequel sont présentés les synonymes est défini selon un calcul de proximité basé sur le principe suivant :

> [S]i un synonyme recouvre beaucoup de sens élémentaires du mot-vedette, il est assez proche de ce dernier au point de vue sémantique. (Manguin *et al.*, 2004, p. 6)

Il s'agit donc de s'appuyer sur le réseau que forment les synonymes pour faire apparaître en début de liste ceux qui partagent le plus de cliques avec le mot vedette. Autrement dit, le site fournit en premier lieu les synonymes les plus probables.

Ce mode de présentation pose un problème majeur : l'ordre dans lequel sont présentés les synonymes est statique, dans le sens où il ne prend pas en compte le contexte dans lequel se trouve le mot dont l'utilisateur cherche un synonyme. Sans cette information, il est impossible de prédire quels synonymes répondront le mieux au besoin de l'utilisateur. Or, on sait que tous les synonymes d'un mot ne sont pas également adaptés à un contexte donné (Murphy, 2003) : par exemple, *repli* sera un meilleur synonyme de *retrait* dans un texte d'actualité qui évoque la progression d'une force armée alors que *reflux* sera plus adapté à un texte relatif au domaine maritime, de la même façon qu'*abolition* sera plus pertinent dans un texte de loi ou *amputation* dans un texte relatif à la médecine. Le fait de mettre en avant les synonymes les plus probables constitue donc un pis-aller face à l'absence d'informations sur le contexte d'apparition du mot qui a été donné en requête (ou du contexte dans lequel va s'insérer le synonyme).

L'AD permet de synthétiser les contextes d'apparition d'un mot dans un corpus donné et de rapprocher les mots qui partagent ces contextes. Nous proposons donc ici d'utiliser cette méthode pour filtrer dynamiquement les synonymes proposés par le DES. Ainsi, plutôt que de renvoyer à l'utilisateur les synonymes les plus probables du mot vedette étant donné un réseau de synonymes construit *in abstracto*, il s'agit de renvoyer les synonymes les plus pertinents pour un type de corpus donné. À titre d'illustration, nous avons rapporté ci-dessous les synonymes du nom

1. Ces proportions s'appuient sur un étiquetage automatique accompagné d'une vérification manuelle effectués par Mai Ho-Dac et Franck Sajous (CLLE-ERSS).
2. `http://www.crisco.unicaen.fr/des/`
3. `http//www.cnrtl.fr/synonymie/`

© ATALA

commission. Les mots qui ont été rapprochés de *commission* suite à l'analyse d'un corpus constitué d'articles de Wikipédia apparaissent en gras :

> *prime, courtage, rémunération, gain, salaire, attribution, remise,* ***intérêt, délégation, comité, charge, traitement, réunion, pouvoir****, paiement,* ***mission, mandat, message, titre, groupe, course****, change,* ***bureau, brevet, besoin, tribunal***

On voit ainsi que les mots *prime*, *rémunération*, *gain*, *salaire* et *paiement*, bien qu'ils apparaissent dans le corpus, ne partagent pas suffisamment de contextes d'apparition avec *commission* pour être captés comme des voisins distributionnels. Or, ces mots relèvent d'une acception de *commission* bien particulière, celle de "Rétribution perçue par le commissionnaire" [4]. On peut donc en déduire que dans un corpus constitué d'articles de Wikipédia, le mot *commission* n'est pas – ou est peu – employé dans des contextes où il renvoie à une valeur pécuniaire. En revanche, il partage les mêmes contextes que :

- *délégation*, *comité*, *groupe* ou *bureau*, qui renvoient à l'acception "Ensemble de personnes officiellement chargées d'une mission à caractère public",
- *charge*, *message*, *mission* ou *mandat*, qui relèvent du sens "Charge qu'une personne reçoit de faire quelque chose".

On se base donc ici sur une AD pour extraire les synonymes qui renvoient au mot vedette tel qu'il est employé dans le corpus.

3 Les voisins distributionnels

Nous appelons *voisins distributionnels* les ressources distributionnelles générées au sein du laboratoire CLLE-ERSS par la chaîne de traitement Syntex-Upéry.

Développé par Didier Bourigault (2007), Syntex est un analyseur syntaxique en dépendance. Les analyses qu'il produit servent d'entrée à Upéry (Bourigault, 2002), qui est le module qui procède au calcul des voisins. Elles sont dans un premier temps ramenées sous la forme de triplets syntaxiques <*mot1*, RELATION, *mot2*>. Par exemple, on obtient les triplets suivants à partir de l'analyse de la phrase *Les paysans révolutionnaires détruisirent le couvent* (seuls les triplets dont *mot1* et *mot2* sont des noms, des verbes, des adjectifs ou des syntagmes nominaux sont retenus) :

- *détruire*, SUJ, *paysan*
- *détruire*, OBJ, *couvent*
- *paysan*, MOD, *révolutionnaire*

Pour les besoins du calcul distributionnel, les triplets sont d'abord réduits sous la forme de couples de type <*mot1*, *mot2*_REL>. Dans la terminologie d'Upéry, on parle de prédicats et d'arguments :

- le prédicat résulte de la concaténation du recteur et de l'étiquette de la relation,
- l'argument correspond au mot qui est régi.

Les résultats de la conversion des exemples précédents au format <*argument*, *prédicat*> sont donc les suivants :

- <*paysan*, *détruire*_SUJ>
- <*couvent*, *détruire*_OBJ>
- <*révolutionnaire*, *paysan*_MOD>

On obtient ainsi deux types d'entités : des prédicats dont les contextes sont tous les arguments

4. Les définitions que nous donnons sont issues du TLFi (`http://atilf.atilf.fr/`).

© ATALA

avec lesquels ils apparaissent dans le corpus, et *vice versa*. Ce formalisme va donner lieu à des rapprochements distributionnels qui opèrent à deux niveaux :
– les arguments sont rapprochés entre eux en fonction des prédicats qu'ils partagent.

couple d'arguments	prédicats partagés
paysan/ouvrier	*travailler*_SUJ, *recruter*_OBJ...
couvent/temple	*bâtir*_OBJ, *piller*_OBJ...

– réciproquement, les prédicats sont rapprochés entre eux en fonction des arguments partagés.

couple de prédicats	arguments partagés
*détruire*_OBJ/*construire*_OBJ	*immeuble*, *bâtiment*...

Étant donné le filtrage sur les catégories grammaticales qui a été opéré en amont, on peut recenser trois types de prédicats : les prédicats verbaux, nominaux et adjectivaux. Les relations qu'ils peuvent porter sont les relations sujet ou objet (prédicats verbaux), modifieur (prédicats adjectivaux) ou *prep* (tous types de prédicats), qui apparaît quand une expansion prépositionnelle est rattachée au nom, au verbe, ou à l'adjectif. Dans ce cas, c'est la préposition qui est accolée au recteur dans le prédicat (<*canon*, *détruire*_AVEC>).

L'étape suivante consiste à associer à chaque prédicat un vecteur constitué de l'ensemble des arguments qu'il prend dans le corpus (et réciproquement). Ces vecteurs sont ensuite comparés les uns aux autres afin de rapprocher les prédicats/arguments qui ont la similarité distributionnelle la plus élevée. La mesure de similarité qui est utilisée dans les voisins distributionnels est l'indice de Lin (1998). Le score de similarité de deux prédicats ou arguments varie – de 0 à 1 – en fonction de plusieurs facteurs : le nombre de contextes partagés, le nombre de triplets différents dans lesquels chacun de deux mots apparaît (indice de productivité), le degré de spécificité du contexte qui permet d'effectuer le rapprochement (se reporter à Bourigault (2002) pour les détails de la procédure de calcul).

Dans le cadre de cette étude, nous utilisons trois ressources distributionnelles qui ont été produites par la chaîne Syntex-Upéry à partir de trois corpus de différentes natures :
– un corpus d'environ 262 millions de mots constitué de l'intégralité des articles de l'encyclopédie en ligne Wikipédia [5] dans sa version de juin 2008,
– un corpus comprenant l'ensemble des articles parus dans le journal Le Monde sur une période de 10 ans – de 1991 à 2000 – soit environ 200 millions de mots,
– un corpus d'environ 30 millions de mots constitué de 515 romans datant du XXe siècle issus de la base Frantext [6].

Ces trois corpus ont donc permis de générer les ressources appelées *voisins de Wikipédia* (désormais *VDW*), *voisins de Le Monde* (*VDLM*) et *voisins de Frantext* (*VDF*).

La taille de ces corpus paraît modeste en comparaison de ceux qui ont pu être utilisés dans des travaux comme ceux de Turney (2008) ou Baroni et Lenci (2010) qui utilisent les corpus issus du Web, dont le nombre de mots se compte en milliards. La raison pour laquelle nous avons malgré tout travaillé avec ces ressources est double. D'une part, elles sont les seules actuellement disponibles au laboratoire CLLE-ERSS (qui est en train de se doter de bases distributionnelles

5. http://fr.wikipedia.org/
6. http://www.frantext.fr

© ATALA

calculées à l'aide de l'analyseur *open source* Talismane [7] (Urieli et Tanguy, 2013)). D'autre part, nous nous situons dans une démarche qui s'appuie notamment sur des retours au contexte pour expliquer les phénomènes observés. Or, comme nous l'avons évoqué dans l'introduction, l'utilisation de corpus issus du Web complique ce type d'approche. Les corpus que nous utilisons ont donc l'avantage qu'ils appartiennent à des genres identifiés et connus.

La différence de taille entre les trois corpus utilisés se répercute sur celle des trois bases de voisins. Sont considérés comme des *voisins* tous les couples de prédicats ou d'arguments dont l'indice de Lin est supérieur ou égal à 0,1. On compte ainsi :
– 3 922 657 couples dans les VDW,
– 5 525 480 couples dans les VDLM,
– 792 356 couples dans les VDF.

Ces trois bases sont consultables sur la plate-forme REDAC [8].

4 Sélection des données à analyser

Notre approche consiste à sélectionner des mots vedettes du DES et à analyser, parmi tous les synonymes auxquels ils sont associés dans le lexique, ceux qui sont – ou ne sont pas – captés comme des voisins distributionnels dans les VDW, VDLM et VDF. Il nous sera ainsi possible d'étudier la façon dont les corpus orientent le repérage de certains synonymes.

Si l'on mesure successivement le recouvrement entre ces trois bases de voisins et le DES en ne prenant en compte que le lexique qui est partagé par les ressources comparées, la proportion de synonymes dans les VDW, VDLM et VDF est respectivement de 1,8 %, 1,5 % et 3,5 %. Réciproquement, la proportion de synonymes du DES reconnus comme des voisins distributionnels est de 41,6 % quand la comparaison se fait avec les VDW, de 33,7 % avec les VDLM et de 21,1 % avec les VDF. Ainsi, seulement un tiers des couples de synonymes du DES sont analysés comme des voisins distributionnels dans nos ressources. Cela laisse supposer que certains couples de synonymes ne sont pas suffisamment substituables en corpus pour que l'AD puisse les rapprocher.

Une des raisons à cette non-substituabilité peut être liée à des propriétés statistiques des couples. En effet, Weeds (2003) montre que les mesures de similarité utilisées pour comparer les vecteurs de contextes ont tendance à rapprocher davantage les mots qui ont des fréquences comparables. De ce fait, une certaine proportion des couples de synonymes – ceux qui sont constitués de mots dont les fréquences sont déséquilibrées – ont statistiquement moins de chance d'être captés par l'AD. Par la suite, nous avons cherché à écarter ces couples de nos échantillons : les couples de synonymes que nous avons voulu étudier en priorité sont ceux qui n'étaient pas sujets au biais évoqué par Weeds. Autrement dit, ceux qui avaient toutes les chances d'être identifiés par l'AD mais qui ne l'ont pas été.

Nous avons distingué les couples les plus susceptibles d'être extraits par l'AD de ceux qui sont pénalisés en calculant le rapport entre la productivité – *r_prod* – de leurs deux membres. La productivité, que nous définissons comme le nombre de contextes différents dans lesquels apparaît un mot, nous semble en effet plus représentative de la distribution d'un mot que sa fréquence (ces deux valeurs sont toutefois très liées (Morlane-Hondère et Fabre, 2012)). Par

7. Il est librement téléchargeable à l'adresse suivante : `http://redac.univ-tlse2.fr/applications/talismane.html`
8. `http://redac.univ-tlse2.fr/index.html`

 © ATALA

exemple les synonymes *monarque* et *chef* ont des productivités respectives de 59 et de 1941, leur r_prod est donc de 0,03 (soit le résultat de la division de 59 par 1941). Nous nous intéressons davantage à l'analyse de couples dont les productivités sont moins inégales comme *aspect* et *air*, qui apparaissent respectivement dans 883 et 935 contextes différents, et dont le r_prod est de 0,94.

Nous avons choisi de ne conserver que les 52 092 couples de synonymes dont le r_prod est supérieur ou égal à 0,23. La raison en est que, sur le 112 863 couples du DES dont le lexique est commun à celui des VDW, ceux qui ont un r_prod inférieur à 0,23 sont majoritairement des non-voisins. Au delà, le rapport s'inverse. Cela confirme que les couples composés de deux mots dont les productivités sont inégales (et donc dont le r_prod est faible), ont moins de chance d'être captés par l'AD.

5 Filtrage des synonymes par les VDW

Avec l'exemple de *commission*, commenté à la section 2, nous avons vu qu'un mot vedette n'était pas forcément voisin avec tous ses synonymes. Dans cette première étude, nous isolons les mots vedettes dont les synonymes sont le moins bien repérés afin d'identifier les conditions par lesquelles se fait le filtrage.

Nous avons choisi de nous intéresser aux cas les plus emblématiques du décalage entre le DES et les VDW, à savoir les mots vedettes dont le nombre de synonymes extraits est quasi nul (mais dont le nombre de synonymes total est supérieur ou égal à 10). Nous avons donc calculé la proportion de synonymes captés par mot vedette et extrait les 30 noms, les 30 verbes et les 30 adjectifs dont la proportion est la plus basse, c'est-à-dire inférieure à 0,1 %. Ces derniers constituent notre échantillon d'étude. Nous en avons rapporté un extrait ci-dessous (le chiffre entre parenthèses correspond au nombre de synonymes qui n'ont pas été extraits comme des voisins, soit la quasi totalité d'entre eux) :

– noms : *éclat* (29), *remède* (21), *espérance* (20), *flamme* (20), *pli* (18), *vide* (17), *ardeur* (17), *clôture* (14), *achèvement* (14), *attente* (14)
– verbes : *exciter* (29), *vider* (22), *coller* (22), *ôter* (20), *arracher* (20), *ébranler* (17), *calmer* (17), *altérer* (17), *allonger* (16), *revêtir* (15)
– adjectifs : *fou* (30), *doux* (27), *rude* (27), *ferme* (25), *ardent* (19), *vague* (18), *habile* (16), *brut* (16), *honnête* (16), *épais* (15)

Cette section est consacrée à la description des trois causes de décalage entre le DES et les VDW que l'analyse de notre échantillon nous a permis d'identifier, à savoir la polysémie (5.1), la connotation (5.2) et la dénotation périphérique (5.3). Les exemples que nous analysons sont ceux qui, dans notre échantillon, nous ont paru les plus représentatifs des phénomènes mis au jour.

5.1 La polysémie

Le phénomène de polysémie correspond à l'explication la plus simple du non-repérage de certains synonymes. En effet, quand deux mots sont synonymes, c'est toujours en fonction d'une acception donnée (Lehmann et Martin-Berthet, 2011). Si cette acception ne se manifeste pas – ou peu – dans le corpus, alors les deux mots ne partageront pas suffisamment de contextes pour être extraits par l'AD. Cette situation peut s'observer sur de nombreux mots de notre échantillon

 © ATALA

comme *pli*, *clôture* ou *fraternité* :

– *pli* s'emploie comme objet des verbes *oblitérer*, *affranchir* ou *poster*. Le sens qui émerge est celui de "enveloppe renfermant une lettre", ses premiers voisins sont *levée*, *courrier*, *paquet*. Ses synonymes *arête*, *repli*, *sillon*, *etc.* ont donc peu de chance d'être captés.
– *clôture* s'emploie principalement en expansion de noms d'évènements comme *jour*, *gala* ou *cérémonie*. Ses contextes d'apparition sont incompatibles avec des synonymes comme *muraille*, *grille* ou *barrage*.
– *fraternité* émerge dans le sens de "communauté ou groupement, laïc ou religieux" dans des contextes comme *rejoindre*_OBJ, *membre*_DE ou encore *fondateur*_DE alors que ses synonymes – *charité*, *sympathie*, *confiance*, *etc.* – réfèrent pour la plupart à son acception "sentiment de solidarité et d'amitié".

L'influence du corpus s'observe de façon plus systématique lors de l'étude des mots *complication* ou *anormal*, qui nous a permis d'identifier un phénomène récurrent lié à la présence de textes relevant du domaine médical dans le corpus Wikipédia. En effet, dans le corpus Wikipédia, ces mots sont utilisés comme des termes médicaux. Ils apparaissent donc dans des contextes relevant du domaine de la médecine, ce qui bloque le repérage de leurs synonymes, qui réfèrent à des acceptions non spécialisées. Par exemple, *complication* apparaît dans des contextes comme *dépister*_OBJ, *mourir*_DE, *prévenir*_OBJ et est modifié par *neurologique*, *cardio-vasculaire* ou *hépatique*. Cela montre clairement que *complication* adopte ici le sens de "phénomènes pathologiques nouveaux résultant de l'évolution d'une maladie et appelant généralement un traitement particulier". On voit donc aisément pourquoi les synonymes que renvoie le DES pour *complication*, à savoir *piège*, *chaos*, *labyrinthe* ou *combinaison* ne sont pas captés comme des voisins : les contextes dans lesquels *complication* apparaît sont tellement spécifiques que sa substitution par l'un de ses synonymes est inenvisageable. On peut également citer le cas de l'adjectif *anormal*, qui modifie principalement des noms comme *hémoglobine*, *saignement*, *gène* ou *protéine*, qui sont des contextes incompatibles avec des synonymes comme *bizarre*, *paradoxal* ou *insolite*. Ainsi, le fait que le corpus Wikipédia contienne une certaine proportion de textes relevant du domaine médical a une influence sur le sens de ces mots, qui prennent alors un sens spécialisé. En conséquence, leur distribution est radicalement différente de celles de leurs synonymes dans le DES.

Nous avons également considéré comme un phénomène de polysémie le cas des emplois métaphoriques. Ce cas de figure s'observe sur un mot vedette comme *flamme*. Le nom *flamme* apparaît en effet dans des contextes comme *lécher*_SUJ, *brûler*_AVEC ou *cerner*_SUJ, c'est-à-dire dans son acception "Mélange gazeux en combustion, dégageant de la chaleur et généralement de la lumière, produit par une matière qui brûle". Ici, le décalage naît du fait que la plupart de ses synonymes – *passion*, *désir*, *enthousiasme*, *etc.* – renvoient à un sens métaphorique de *flamme*. Ces derniers partagent dans le corpus Wikipédia des contextes comme *manifester*_OBJ, *susciter*_OBJ ou *provoquer*_OBJ, dans lesquels *flamme* n'apparaît pas.

5.2 La connotation

Nous avons observé dans notre échantillon deux types de connotations que Kerbrat-Orecchioni (1977) désigne sous le nom de "connotation énonciative" et "connotation stylistique".

Dans le premier cas, le décalage naît du fait que le mot vedette ou son synonyme porte une valeur axiologique que l'autre n'a pas. Ce phénomène est dit "énonciatif" dans le sens où l'emploi d'un mot axiologiquement marqué plutôt que d'un mot *neutre* donne des indications "non sur le référent du message, mais sur son énonciateur" (Kerbrat-Orecchioni, 1977, p. 104). Les mots

 © ATALA

marqués et non marqués partagent le même sens dénotatif, ils peuvent donc tout de même constituer des synonymes potentiels (Murphy, 2003). Parmi les mots vedettes de notre échantillon, les adjectifs *sommaire* et *fou* nous permettent d'illustrer ce phénomène. Ils ont respectivement pour synonymes *simpliste* et *stupide*, lesquels sont clairement péjoratifs. Ils n'ont pas été extraits par l'AD. Si l'on regarde de plus près le cas de *fou*/*stupide*, on peut voir que les premiers voisins de *fou* sont *malade*, *immortel* et *endormi* et que ceux de *stupide* sont *intelligent* et *curieux*. Cela confirme le fait que *fou* est utilisé comme un adjectif non marqué alors que *stupide* est rapproché d'adjectifs qui portent une valeur subjective. Il est à noter que, puisque la rédaction des articles de Wikipédia est soumise au respect de la neutralité de point de vue [9], qui proscrit l'emploi de mots connotés, un mot comme *stupide* apparaît majoritairement dans des citations ou des titres d'œuvres. De ce fait, la spécificité de ces contextes peut créer un décalage distributionnel entre les synonymes axiologiquement marqués et leurs équivalents non marqués. Ces cas de décalage sont ainsi révélateurs d'un certain degré d'hétérogénéité du corpus Wikipédia.

Le deuxième type de connotation – la connotation stylistique – renvoie au registre de langue auquel appartient un mot. Nos données montrent que si deux mots qui appartiennent à des registres différents peuvent présenter des "propriétés sémiotiques" identiques, cela n'implique pas qu'ils seront substituables dans tous les contextes. En effet, certains mots de la langue *standard* peuvent prendre des emplois argotiques, comme c'est le cas de *veine*, qui figure dans le DES comme un synonyme de *hasard*. Or, *veine* s'emploie dans le corpus dans des contextes comme *couler*_DANS, *injection*_DANS, *occlusion*_DE ou peut être modifié par les adjectifs *splénique*, *fémoral*, *jugulaire*, *etc.* Ces contextes nous montrent clairement que *veine* est employé dans le corpus comme un type de vaisseau sanguin et non au sens de "Chance, fortune". Il n'apparaît pas dans les même contextes que *hasard* comme *part*_DE, *jeu*_DE, *rencontrer*_PAR, *etc.* On peut tirer les mêmes conclusions pour le mot vedette *chauffer*, qui, dans le DES, a pour synonymes *voler* et *dérober*, qui renvoient à une acception argotique du mot.

5.3 La dénotation périphérique

À la suite de Murphy (2003), nous distinguons dans le sens dénotatif d'un mot son noyau sémantique (*core meaning*) et ses – éventuels – traits périphériques (*peripheral features*). La notion de trait périphérique nous permet en effet d'aborder le problème des synonymes qui partagent un même noyau de sens (ce qui n'était pas le cas dans la section 5.1 : un *pli* (postal) n'est pas la même chose qu'un *repli*) mais qui se distinguent du point de vue des nuances sémantiques qu'ils peuvent véhiculer (qui sont donc exprimées par ces traits).

Parmi les mots vedettes de notre échantillon, le verbe *vider* est représentatif de ce phénomène. Les synonymes qui sont listés par le DES pour *vider* apportent des précisions sur l'action exprimée par le verbe :
- *épuiser* : "Vider (quelque chose) de son contenu ou de sa substance",
- *écoper* : "Vider l'eau qui s'accumule au fond d'une embarcation non pontée, à l'aide d'une écope",
- *évacuer* : "Vider (un pays, un lieu) des personnes qui l'occupent",
- *déménager* : "Vider (le meuble, la pièce) de tout ce qu'il contient",
- *assécher* : "Vider de ses ressources".

Les traits véhiculés par ces synonymes apportent des précisions soit sur la nature des compléments

9. `http://fr.wikipedia.org/wiki/Wikipédia:Neutralité_de_point_de_vue`

 © ATALA

*vider*_OBJ	*épuiser*_OBJ	*évacuer*_OBJ	*déménager*_OBJ	*assécher*_OBJ
chargeur	deux édition	ville	siège social	marécage
poubelle	pioche	eau	franchise	marais
cuve	gisement	habitant	local	étang
récipient	stock	chaleur	colonel	lac
caisse	munition	population	fois	rivière
coffre	réserve	personne	rue	air
cache	recours	partie	siège	zone
sac	carburant	troupe	capitale	terre
querelle	ressource	territoire	mois	partie
cave	combustible	île	centre	

TABLE 1 – Différences dans la distribution de *vider* et de quelques-uns de ses synonymes.

du verbe, soit sur la manière dont s'effectue l'action. On peut supposer que ces nuances se répercutent sur la distribution de ces synonymes. Ainsi, afin d'observer ce phénomène, nous avons rapporté au tableau 1 les dix arguments [10] qui ont l'information mutuelle la plus élevée avec *vider* et ses synonymes *épuiser*, *évacuer*, *déménager* et *assécher* lorsqu'ils portent la fonction OBJ [11].

Les contextes recensés dans le tableau 1 confirment l'hypothèse que nous avons formulée plus haut, à savoir que les traits portés par les synonymes de *vider* ont une influence sur leurs distributions dans le corpus. Ainsi, alors que *vider* prend pour objets des noms de contenants (*récipient*, *caisse*, *coffre*), *épuiser* s'emploie avec des noms de ressources (*carburant*, *combustible*... *ressource*) ou de *lieux* contenant ces ressources (*stock*, *réserve*, *gisement*), *évacuer* avec des noms de lieux (*ville*, *territoire*, *île*) ou d'animés (*habitant*, *population*, *personne*), *déménager* avec des noms de lieux (*local*, *rue*, *capitale*) et *assécher* avec des noms de lieux qui contiennent habituellement de l'eau (*marécage*, *étang*, *lac*). De ce fait, alors que *vider* et ses synonymes partagent un même noyau de sens, on s'aperçoit que le chevauchement entre les noms qu'ils prennent comme objets reste très faible. Ils semblent donc davantage fonctionner sur le mode de la complémentarité plutôt que sur celui de la substituabilité : *évacuer*, *épuiser*, *déménager* et *assécher* portent des traits périphériques qui font que ces mots s'emploient dans des contextes spécifiques. Le fait que *vider* ne porte pas de trait périphérique semble bloquer son apparition dans ces contextes, donc sa substituabilité avec ses synonymes.

Les trois cas de figure que nous avons commentés dans cette section 5 nous ont permis de mettre en lumière certains des fonctionnements qui entraînent l'absence d'un synonyme dans une base de voisins distributionnels. En excluant les synonymes qui ne sont pas – ou très peu – substituables avec un mot vedette donné, l'AD permet de faire émerger ceux qui sont le plus susceptibles d'apparaître dans le contexte du mot vedette, autrement dit, les plus pertinents pour un utilisateur du DES. De ce fait, une hypothétique intégration du DES aux pages de modification des articles de Wikipédia gagnerait à proposer en priorité à l'utilisateur les synonymes sélectionnés par les VDW (au détriment des synonymes qui, par exemple, renvoient à des acceptions du mot vedette qui ne se manifestent pas dans le corpus).

10. Sauf pour *assécher*_OBJ, qui n'en compte que 9 au total.

11. Nous n'avons pas fait figurer les contextes du verbe *écoper* étant donné qu'il n'apparaît dans le corpus que dans le sens de "Subir (des dommages matériels), recevoir (des coups) ; être atteint ou touché (par quelque chose que l'on subit)".

© ATALA

6 Variation du repérage en fonction du corpus

Nous avons décrit, dans la section précédente, quelques unes des propriétés qui régissent la substituabilité des mots vedettes et de leurs synonymes. Dans cette section, nous montrons que la nature des synonymes qui sont rapprochés du mot vedette par l'AD varie selon que la base distributionnelle utilisée ait été calculée à partir de textes encyclopédiques, journalistiques, ou littéraires. Nous avons donc, dans un premier temps, comparé successivement la proportion de synonymes repérés comme des voisins par les VDW avec celle des VDLM et des VDF.

	Synonymes voisins	Synonymes voisins partagés avec les VDW
VDLM	11,4	10
VDF	2,5	2,1

TABLE 2 – Nombres moyens de synonymes repérés dans les VDLM et les VDF.

Les résultats, rapportés au tableau 2, montrent qu'en moyenne, les VDLM extraient 11,4 synonymes par mot vedette (sur 16), ce qui correspond exactement au nombre de synonymes captés par les VDW (la différence dans la proportion de synonymes extraits évoquée à la section 4 ne s'observe donc pas sur les données seuillées). En revanche, les VDF n'en extraient que 2,5. Nous attribuons la faiblesse de ce chiffre à la taille du corpus Frantext. On voit également que la plupart des synonymes captés par les VDLM/VDF l'ont également été par les VDW. Cela montre qu'en moyenne, la plupart des synonymes qui sont extraits par les VDLM et les VDF le sont aussi par les VDW. On note que la réciproque n'est pas vraie pour les VDF puisque les VDW reconnaissent en moyenne 4,6 fois plus de synonymes comme des voisins distributionnels que les VDF. Le fait qu'on observe un recouvrement important entre les VDW et les VDLM montre que les fonctionnements des mots vedettes et de leurs synonymes dans les corpus Wikipédia et Le Monde ne sont pas radicalement différents. Toutefois, ces résultats globaux ne doivent pas masquer le fait que les effets du corpus se font malgré tout ressentir de façon importante pour une certaine proportion de mots vedettes. À titre d'exemple, 16 % des 1202 mots vedettes observés possèdent au moins 5 synonymes qui n'ont été captés que par les VDW ou les VDLM. C'est notamment le cas du mot vedette *ton*, que nous décrivons ci-dessous.

Nous avons rapporté dans le tableau 3 la liste des synonymes du nom *ton* selon qu'ils sont captés par les bases de voisins (✓), qu'ils sont présents dans leur lexique mais non repérés comme des voisins (✗) ou absents de leur lexique (∅). Ce tableau appelle plusieurs observations. Tout d'abord, on constate que, sur 34 synonymes, beaucoup sont absents des lexiques des trois ressources. Cela est dû à la rigueur du filtrage opéré en amont. Ainsi, le mot *main*, qui apparaît dans la version non filtrée des VDW, est absent de la version seuillée (avec un r_prod à 0,23), qui ne contient que les couples les plus susceptibles d'être captés. Il est à noter que nous n'avons pas fait apparaître les synonymes qui n'apparaissent dans aucun des lexiques des trois ressources.

Le troisième point est le plus important pour notre problématique. En effet, on remarque que même si les listes de synonymes captés par les VDW et les VDLM sont de tailles à peu près identiques, leur contenu est relativement différent :

– moins de la moitié des synonymes reconnus respectivement par les VDW et les VDLM sont communs aux deux ressources. Ces synonymes sont les suivants : *accent*, *écriture*, *goût*, *manière*, *son*. Ces mots partagent donc les mêmes contextes d'apparition que *ton* que ce soit dans le

 © ATALA

synonyme	VDW	VDLM	VDF	synonyme	VDW	VDLM	VDF
accent	✓	✓	✓	note	✓	✗	✗
air		✓	✓	nuance	✓	✗	
bruit	✗	✗	✗	parole	✓	✗	✗
corde	✗		✗	patte	✓		✗
couleur		✓	✗	plume	✓	✗	✗
écho	✗	✓		procédé	✗	✗	
écriture	✓	✓		signature	✓	✗	
expression		✓	✓	son	✓	✓	✗
façon	✓	✓	✗	style		✓	
facture	✗	✗		teinte	✓		
genre		✓	✗	tension	✗	✗	
goût	✓	✓	✗	timbre	✗	✗	
griffe	✗			tonalité	✓		
main		✗	✗	touche	✗	✗	
manière	✓	✓	✗	tour		✗	✗
mode		✗		verbe	✗		
musique		✗	✗	voix		✓	✓

TABLE 3 – Repérage des synonymes du nom *ton* en fonction des bases de voisins.

corpus Wikipédia ou dans le corpus Le Monde.

– a l'inverse, les mots *nuance*, *signature*, *parole*, *plume* et *note* ne partagent les mêmes contextes que *ton* que dans le corpus Wikipédia. On peut supposer que ces rapprochements sont à attribuer à une plus grande présence du vocabulaire des arts dans les VDW. Une comparaison des modifieurs les plus fréquents de *ton* dans les corpus Wikipédia et Le Monde va dans le sens de cette intuition : alors que *ton* est fréquemment modifié par des adjectifs comme *majeur*, *mineur*, *chaud*, *bleu* ou *clair* dans le corpus Wikipédia, on trouve parmi ses modifieurs privilégiés dans le corpus Le Monde des adjectifs comme *grave*, *ferme*, *vif*, *modéré* ou *solennel*. Dans le deuxième cas, l'emploi de *ton* privilégié semble donc être le suivant : "Qualité de la voix (hauteur, timbre, intensité)".

On a ici pu déduire de l'analyse des synonymes de *ton* captés par les différentes bases de voisins que le sens dans lequel était employé ce mot différait en fonction de la base distributionnelle considérée. Cette observation traduit une différence d'usage dans les corpus qui ont permis de générer ces ressources distributionnelles.

À ce stade, il est également possible de mener une comparaison des synonymes captés par les différentes bases de voisins qui va au delà de la dichotomie absence/présence. En effet, même si dans la plupart des cas, les synonymes extraits par les VDW et les VDLM pour un mot vedette donné sont à peu près les mêmes, il n'est pas évident que ces synonymes aient un score de similarité distributionnelle avec le mot vedette identique dans les deux ressources. À titre d'exemple, les 23 synonymes du mot vedette *tour* ont tous été captés par les VDW et les VDLM. Toutefois, on peut voir ci-dessous que les cinq meilleurs voisins de *tour* portent sur deux acceptions différentes en fonction de la ressource :

– VDW : *bâtiment* (0,47), *construction* (0,345), *pièce* (0,257), *course* (0,186), *forme* (0,166)
– VDLM : *course* (0,217), *tournée* (0,197), *façon* (0,190), *voyage* (0,170), *marche* (0,169)

De ce fait, il devient possible d'observer les effets du corpus à un niveau de finesse plus élevé que

© ATALA

celui que nous avons adopté plus haut lors de l'étude du repérage des synonymes de *ton*.

7 Conclusion

Dans cette étude, nous avons abordé la question de ce que pourrait apporter à un dictionnaire de synonymes un filtrage par une base distributionnelle. Cela nous a amené à montrer, dans un premier temps, que tous les synonymes d'un mot vedette donné ne partagent pas ses contextes d'apparition. Des phénomènes comme la polysémie, la connotation ou des incompatibilités en termes de traits dénotatifs font que la distribution du mot vedette et de certains de ses synonymes divergent. De ce fait, l'AD permet d'opérer une sélection parmi les synonymes proposés pour un mot vedette donné.

Nous avons ensuite montré que ce phénomène pouvait varier selon que l'on filtrait le dictionnaire avec une ressource calculée à partir d'un corpus constitué de textes encyclopédiques, journalistiques ou littéraires. La pertinence des synonymes d'un mot vedette donné n'est en effet pas absolue mais varie en fonction du type de corpus dans lequel le mot est employé.

Nous proposons par la suite d'exploiter les informations que fournit le corpus sur la distribution du mot dont on recherche le synonyme. Il s'agirait alors de se servir de ressources distributionnelles en support d'un dictionnaire de synonymes pour savoir quels sont ceux qui sont les plus pertinents pour un type de texte donné et de réorganiser les résultats fournis en conséquence. Une étape d'évaluation serait nécessaire pour mesurer les apports d'une telle démarche. Nous avons donc prévu de poursuivre la présente étude en mettant en place un protocole d'évaluation dans lequel il sera demandé à des utilisateurs de sélectionner les meilleurs synonymes pour un mot donné dans un ensemble de phrases extraites des corpus Wikipédia, Le Monde et Frantext. Nous faisons ainsi l'hypothèse que le choix des utilisateurs se portera en priorité sur les synonymes qui sont voisins du mot cible dans la base distributionnelle calculée à partir du corpus d'où a été extraite la phrase.

Références

BARONI, M. et LENCI, A. (2010). Distributional Memory : A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36(4):673–721.

BARONI, M. et LENCI, A. (2011). How we BLESSed distributional semantic evaluation. *In Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS'11, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

BOURIGAULT, D. (2002). UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *In Actes de la 9e conférence sur le Traitement Automatique de la Langue Naturelle* (24–27 juin 2002), Nancy.

BOURIGAULT, D. (2007). *Un analyseur syntaxique opérationnel :* SYNTEX. Habilitation à diriger des recherches. Université Toulouse II – Le Mirail.

BOURIGAULT, D. et GALY, E. (2005). Analyse distributionnelle de corpus de langue générale et synonymie. *In 4e journées de la linguistique de corpus* (15–17 septembre 2005), Lorient.

GREFENSTETTE, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA, USA.

 © ATALA

HABERT, B. (1998). *Des mots complexes possibles aux mots complexes existants : l'apport des corpus*. Habilitation à diriger des recherches en linguistique, Université Lille III – Charles de Gaulle.

HARRIS, Z. (1968). Mathematical structures of language. John Wiley & Sons.

KERBRAT-ORECCHIONI, C. (1977). *La connotation*. Linguistique et sémiologie. Presses universitaires de Lyon.

LEHMANN, A. et MARTIN-BERTHET, F. (2011). *Introduction à la lexicologie : sémantique et morphologie*. Collection Lettres supérieures. Armand Colin.

LIN, D. (1998). An information-theoretic definition of similarity. *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304.

MANGUIN, J.-L. (2005). Les dictionnaires en ligne : nouvelles diffusions, nouveaux objectifs. Paris, France.

MANGUIN, J. L., FRANÇOIS, J., EUFE, R., FESENMEIER, L., OZOUF, C. et SÈNÈCHAL, M. (2004). Le dictionnaire électronique des synonymes du CRISCO : un mode d'emploi à trois niveaux. *In Cahiers du CRISCO*, volume 34. CRISCO, Université de Caen.

MORLANE-HONDÈRE, F. et FABRE, C. (2010). L'antonymie observée avec des méthodes de TAL : une relation à la fois syntagmatique et paradigmatique ? *In* ASSOCIATION POUR LE TRAITEMENT AUTOMATIQUE DES LANGUES (ATALA), éditeur : *Actes de TALN 2010*, page 6, Montréal, Canada. Article court.

MORLANE-HONDÈRE, F. et FABRE, C. (2012). Étude des manifestations de la relation de méronymie dans une ressource distributionnelle (Study of Meronymy in a Distribution-Based Lexical Resource) [in French]. *In Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, pages 169–182, Grenoble, France. ATALA/AFCP.

MORLANE-HONDÈRE, F. et FABRE, C. (2012). Le test de substituabilité à l'épreuve des corpus : utiliser l'analyse distributionnelle automatique pour l'étude des relations lexicales. *SHS Web of Conferences*, 1:1001–1015.

MURPHY, M. L. (2003). *Semantic Relations and the Lexicon : Antonymy, Synonymy and other Paradigms*. University Press, Cambridge.

OROSEMANE, L. (2012). Qui se cache derrière le dictionnaire des synonymes de Caen ? *Rue 89*. En ligne. `http://www.rue89.com/2012/10/28/qui-se-cache-derriere-le-dictionnaire-des-synonymes-de-caen-236552` Page consultée le 25/02/2013.

THOT CURSUS (2012). Un dictionnaire des synonymes gros comme ça ! En ligne. `http://cursus.edu/dossiers-articles/articles/8902/dictionnaire-des-synonymes-gros-comme/` Page consultée le 29/03/2013.

TURNEY, P. D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. *In COLING*, pages 905–912.

URIELI, A. et TANGUY, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane (à paraître). *In Actes de TALN 2013*.

WEEDS, J. (2003). *Measures and applications of lexical distributional similarity*. Thèse de doctorat, Université du Sussex.

 © ATALA

Exploration sémantique visuelle à partir des cooccurrences de deuxième et troisième ordre

Ann Bertels[1,2] Dirk Speelman[2]
(1) ILT, University of Leuven, Dekenstraat 6, B-3000 Leuven (Belgique)
(2) QLVL, University of Leuven, Blijde-Inkomststraat 21, B-3000 Leuven (Belgique)
ann.bertels@ilt.kuleuven.be, dirk.speelman@arts.kuleuven.be

RÉSUMÉ

Cet article présente la méthodologie adoptée pour l'exploration visuelle de la proximité sémantique entre les cooccurrents d'un mot-pôle dans un corpus technique. Nous procéderons à des analyses de positionnement multidimensionnel dans le but de compléter des analyses sémantiques quantitatives et de mieux comprendre les phénomènes d'homogénéité et d'hétérogénéité sémantiques.

Nous discuterons les aspects méthodologiques pour le regroupement et la visualisation des cooccurrents de premier ordre d'un mot-pôle technique hétérogène (*tour*). Le regroupement de ces cooccurrents se fait en fonction des cooccurrents de deuxième et troisième ordre qu'ils partagent. Les questions méthodologiques les plus importantes portent sur les mesures d'association, les métriques de distance pour la matrice de dissimilarité et les seuils inférieurs et supérieurs. Elles permettront de trouver la meilleure configuration de paramètres, la plus fiable et la plus efficace, pour l'interprétation sémantique.

ABSTRACT

Visual semantic exploration on the basis of second and third order co-occurrences

This article addresses the methodology for visual exploration of the semantic proximity between the cooccurrents of a node in a technical corpus. Multidimensional scaling analysis is carried out in order to refine quantitative semantic analysis and to come to a better understanding of the phenomena of semantic homogeneity and heterogeneity.

We will discuss the methodological aspects for clustering and plotting the first order co-occurrents of a heterogeneous technical node (*tour*). These co-occurrents are clustered with respect to the second and third order co-occurrents they have in common. Important methodological issues involve association measures, distance measures for the dissimilarity matrix and lower and upper thresholds. They help us find the best reliable and efficient configuration of parameters in support of clear semantic interpretation.

MOTS-CLÉS : Analyse de cooccurrences, cooccurrents de deuxième ordre et de troisième ordre, mesures d'association, métriques de distance, positionnement multidimensionnel.
KEYWORDS : Co-occurrence analysis, second order and third order co-occurrents, association measures, distance measures, MultiDimensional Scaling (MDS).

1 Introduction

Cet article s'inscrit dans la suite d'une étude sémantique quantitative, effectuée sur un corpus spécialisé relevant du domaine technique restreint des machines-outils pour l'usinage des métaux (1,7 million d'occurrences). Le but de cette étude était de vérifier si et

 © ATALA

dans quelle mesure les unités lexicales spécifiques de ce corpus technique sont monosémiques. Cela revenait à étudier la corrélation entre le rang de spécificité et le rang de monosémie de quelque 5000 unités lexicales. À cet effet, nous avions développé une mesure de monosémie, permettant de calculer le degré de monosémie des unités lexicales spécifiques en fonction des cooccurrents de deuxième ordre partagés. Les résultats de l'analyse de régression ont confirmé les observations d'études sémantiques antérieures (Temmerman, 2000 ; Eriksen, 2002 ; Ferrari, 2002). Ils ont aussi permis de réfuter la thèse monosémiste traditionnelle (Wüster, 1991), puisque les unités lexicales simples les plus spécifiques n'étaient pas les plus homogènes sémantiquement (voir Bertels et al., 2010).

Notre mesure de monosémie repose sur l'analyse des cooccurrences (Grossmann et Tutin, 2003 ; Blumenthal et Hausmann, 2006) et envisage la monosémie en termes d'homogénéité sémantique (Habert et al., 2005). Une unité lexicale monosémique apparaît dans des contextes sémantiquement plutôt homogènes et se caractérise par des cooccurrents qui appartiennent à des champs sémantiques similaires. La similarité distributionnelle reflète la similarité sémantique. Par contre, une unité lexicale polysémique se caractérise par des cooccurrents sémantiquement plus hétérogènes. L'accès à la sémantique des cooccurrents d'un mot-pôle se fait à partir de leurs cooccurrents, c'est-à-dire à partir des cooccurrents de deuxième ordre (Grefenstette, 1994). Ceux-ci permettent aussi de remédier au problème de la distribution irrégulière des sens du mot-pôle dans un corpus (Habert et al., 2004). Si les cooccurrents de premier ordre d'un mot-pôle partagent beaucoup de cooccurrents de deuxième ordre, ces derniers se recoupent formellement, ce qui est une indication de l'homogénéité sémantique des cooccurrents de premier ordre et, dès lors, du mot-pôle. Notre mesure de monosémie s'appuie sur le degré de recoupement. Elle a été concrétisée à l'aide de scripts en Python. Pour déterminer les cooccurrents statistiquement pertinents, elle s'appuie sur la mesure statistique du *Log-Likelihood Ratio* (rapport de vraisemblance) (Dunning, 1993). Au seuil de significativité très sévère (valeur $p \leq 0{,}0001$), elle a permis de répertorier uniquement les cooccurrents sémantiquement pertinents. Les résultats de la mesure de monosémie ont été validés à partir d'une analyse manuelle des cooccurrents les plus pertinents et au moyen de plusieurs dictionnaires spécialisés (Bertels et al., 2010).

Notre mesure de monosémie est déjà capable de détecter l'homogénéité sémantique des unités lexicales monosémiques ainsi que l'hétérogénéité sémantique des unités lexicales polysémiques, mais nous n'arrivons pas encore à dissocier les sens polysémiques, ni à faire la distinction entre la polysémie et le vague. À cet effet, nous procédons à des analyses de positionnement multidimensionnel. Le but de ces analyses est de regrouper les cooccurrents de premier ordre d'un mot-pôle et de les visualiser en 2D, en fonction des cooccurrents de deuxième ordre. L'exploration visuelle des proximités sémantiques entre les cooccurrents de premier ordre devrait permettre de mieux interpréter les phénomènes d'homogénéité et d'hétérogénéité sémantiques. Dans le développement qui suit nous décrirons les premières expérimentations pour un mot-pôle hétérogène sémantiquement dans le corpus technique.

Le mot-pôle *tour* se caractérise par un degré de recoupement ou de monosémie plutôt faible. *Tour* a en effet plusieurs sens techniques (« machine-outil pour l'usinage des pièces » et « rotation »), comme l'illustrent les phrases-exemples 1 et 2 ci-dessous. *Tour* possède aussi plusieurs sens généraux, notamment dans les expressions *tour d'horizon* et *à tour de rôle*.

1. *Sur ce tour, on usine l'alésage central.*
2. *Les broches peuvent monter jusqu'à quinze mille tours par minute.*

 © ATALA

Dans cet article, nous discuterons d'abord les aspects méthodologiques à prendre en considération (section 2), d'une part, pour le repérage des cooccurrents et, d'autre part, pour le regroupement (*clustering*) et la visualisation. Dans la 3e section, nous présenterons les résultats des analyses de positionnement multidimensionnel et des représentations graphiques. Ils permettront de trouver la meilleure configuration de paramètres pour l'interprétation des relations entre les cooccurrents de premier ordre et pour l'interprétation sémantique du mot-pôle. Nous terminerons par une conclusion et quelques pistes pour des recherches futures (section 4).

2 Aspects méthodologiques

La plupart des analyses en sémantique distributionnelle (Sahlgren, 2006 ; Peirsman et Geeraerts, 2009 ; Ferret, 2010) déterminent la proximité sémantique entre mots à partir des contextes qu'ils partagent dans un corpus donné, souvent sous forme d'espaces de mots (Turney et Pantel, 2010). Deux mots sont sémantiquement similaires s'ils figurent dans des contextes similaires ou s'ils partagent des cooccurrents de premier ordre. Les analyses en sémantique distributionnelle s'appuient sur des mesures d'association pour déterminer les cooccurrents statistiquement pertinents et sur des métriques de distance pour positionner les mots les uns par rapport aux autres en fonction des cooccurrents de premier ordre qu'ils partagent. Les mots qui apparaissent souvent avec les mêmes cooccurrents se retrouvent regroupés dans un espace de mots. La représentation graphique de cet espace de mots permet de visualiser des groupes de synonymes (Ferret, 2010) ou des groupes de mots sémantiquement liés (Peirsman et Geeraerts, 2009). Si les données au niveau des cooccurrents de premier ordre sont rares (*data sparseness*), il est fait appel aux cooccurrents de deuxième ordre (Lemaire et Denhière, 2006).

Les mesures d'association et les métriques de distance seront également abordées dans cet article, à cette différence près qu'il s'agit ici de la proximité sémantique entre cooccurrents de premier ordre et non entre mots. L'objet d'analyse se situe donc à un ordre supérieur par rapport à l'objet d'analyse des études en sémantique distributionnelle évoquées ci-dessus. L'objectif de notre étude est de positionner les cooccurrents de premier ordre d'un mot-pôle les uns par rapport aux autres en fonction des cooccurrents de deuxième ordre partagés, pour ainsi trouver des groupes de cooccurrents de premier ordre sémantiquement liés. Les cooccurrents de deuxième ordre sont donc définis comme les cooccurrents de premier ordre par rapport aux cooccurrents de premier ordre du mot-pôle dont nous cherchons à visualiser les sens multiples. Si la visualisation des cooccurrents de premier ordre du mot-pôle montre par exemple deux groupes nettement délimités, le mot-pôle se caractérise par deux sens différents. Pour remédier à un problème de rareté des données au niveau des cooccurrents de deuxième ordre, nous descendons d'un niveau pour faire intervenir leurs cooccurrents, c'est-à-dire les cooccurrents de troisième ordre par rapport au mot-pôle, qui sont les cooccurrents d'ordre deux par rapport aux cooccurrents de premier ordre regroupés et visualisés. L'objectif principal de notre étude est de voir plus clair dans les relations sémantiques entre les cooccurrents de premier ordre d'un mot-pôle dans le but de mieux comprendre le degré d'hétérogénéité sémantique de ce mot-pôle, calculé par la mesure de monosémie. Nous ne visons pas à désambiguïser les différents sens d'un mot polysémique. Le positionnement des cooccurrents de premier ordre devrait nous aider à compléter et enrichir la mesure de monosémie pour ainsi affiner l'analyse sémantique quantitative.

 © ATALA

2.1 Analyse des cooccurrences

Pour identifier les cooccurrents pertinents d'un mot-pôle, ainsi que les cooccurrents de deuxième et troisième ordre, les questions méthodologiques les plus importantes portent sur la mesure d'association, la taille de la fenêtre d'observation et le seuil de significativité.

2.1.1 Mesures d'association : LLR et PMI

Dans notre mesure de monosémie, nous avons utilisé la mesure statistique du *Log-Likelihood Ratio* ou LLR pour calculer le degré d'association entre une unité lexicale spécifique (mot-pôle) et ses cooccurrents de premier et deuxième ordre. La mesure du LLR calcule le degré d'association entre deux mots w_1 et w_2, en comparant les vraisemblances (*likelihoods*) L de l'hypothèse d'indépendance (ou $p = P(w_2|w_1) = P(w_2|\neg w_1)$) et de l'hypothèse de dépendance ($p_1 = P(w_2|w_1) \neq P(w_2|\neg w_1) = p_2$), sous l'assomption d'une distribution binomiale des mots (Dunning, 1993). Dans la formule du LLR, indiquée ci-dessous, le numérateur exprime la vraisemblance des données, étant donné une estimation du maximum de vraisemblance pour p (sous l'hypothèse d'indépendance). Le dénominateur exprime la vraisemblance des données, étant donné une estimation du maximum de vraisemblance pour p_1 et p_2 (sous l'hypothèse de dépendance).

$$LLR = -2\, log \frac{L(P(w_2|w_1); p) * L(P(w_2|\neg w_1); p)}{L(P(w_2|w_1); p_1) * L(P(w_2|\neg w_1); p_2)}$$

La mesure du LLR est couramment utilisée comme mesure d'association dans les analyses de cooccurrences. Toutefois, elle souffre d'un biais de fréquence (Peirsman et Geeraerts, 2009). Pour des fréquences (très) élevées, elle gonfle le degré d'association (LLR) de deux mots (très) fréquents, ce qui affecte indirectement le résultat de la mesure de monosémie.

Une autre mesure d'association, souvent utilisée en recherche d'information, est la mesure de l'information mutuelle (*Pointwise Mutual Information* ou PMI) (Church et Hanks, 1990). La mesure de la PMI compare la probabilité de la cooccurrence de deux mots w_1 et w_2 (probabilité conjointe) à la probabilité d'une association simplement due au hasard.

$$PMI(w_1, w_2) = log_2 \frac{P(w_1, w_2)}{P(w_1) * P(w_2)}$$

La mesure de la PMI est plus fiable pour des cooccurrences qui apparaissent souvent ensemble, mais elle est moins fiable pour des cooccurrences dont la co-fréquence est faible, puisqu'elle a tendance à surestimer la valeur d'association des mots rares (Dunning, 1993 ; Evert, 2007). Nous aurons donc tout intérêt à ne considérer que les cooccurrences suffisamment fréquentes et à introduire un seuil inférieur de co-fréquence minimale.

2.1.2 Autres paramètres

L'analyse des cooccurrences soulève également des questions sur la configuration de paramètres la plus stable, en l'occurrence celle qui fournit les informations sémantiques les plus stables et les plus fiables. Pour ce faire, nous avons effectué des expérimentations faisant intervenir trois paramètres, à savoir la forme graphique ou le lemme des cooccurrents de premier et deuxième ordre, la taille de la fenêtre d'observation et le seuil de significativité (cf. Martinez 2005 ; Mayaffre 2008). Ces expérimentations ont permis de trouver la configuration la plus stable (Bertels et Speelman, 2012). Elle faisait intervenir une

 © ATALA

fenêtre d'observation (*span*) de 5 mots à gauche et à droite du mot-pôle et un seuil de significativité très sévère (valeur $p \leq 0{,}0001$) afin de relever uniquement les cooccurrents sémantiquement pertinents. Les cooccurrents de premier et deuxième ordre ont été considérés au niveau des formes graphiques, qui véhiculent des informations sémantiques plus riches, telles que la distinction entre *pièce usinée* (« résultat ») et *pièce à usiner* (« avant le processus d'usinage »). Les mots grammaticaux ont été conservés, étant susceptibles d'apporter des informations sémantiques (par exemple *pendant* indique un processus).

2.1.3 Configuration préconisée

Dans nos expérimentations de regroupement et de visualisation, nous envisageons plusieurs cas de figure afin de comparer les résultats des deux mesures d'association et de les confronter à la mesure de monosémie. Nous considérons les mots-pôles au niveau des lemmes et les cooccurrents de premier, deuxième et troisième ordre au niveau des formes fléchies. La taille de la fenêtre d'observation est préservée (5 mots à gauche et à droite). Pour ces expérimentations, nous travaillons sur un extrait du corpus technique de 320 000 mots. Des scripts en Python permettent d'extraire les informations requises par mot-pôle.

2.2 Regroupement et visualisation

Plusieurs travaux se sont penchés sur le regroupement et la visualisation. Dans un corpus médical, le logiciel Zellig a permis de visualiser des regroupements de mots (au lieu de cooccurrents de premier ordre), en fonction des cooccurrents syntaxiques partagés. Les proximités entre les mots ont été calculées à partir d'arbres d'analyse et elles ont abouti à un découpage en classes (Fabre, Habert et Labbé, 1997). D'autres travaux de visualisation visent à explorer des textes pour accéder au contenu (Roy, 2007).

2.2.1 MDS ou positionnement multidimensionnel

Pour le regroupement (*clustering*) et la visualisation des cooccurrents de premier ordre d'un mot-pôle, en fonction des cooccurrents de deuxième ordre qu'ils partagent, nous recourons à la technique du positionnement multidimensionnel (*MultiDimensional Scaling* ou MDS) (Kruskal et Wish, 1978 ; Cox et Cox, 2001 ; Venables et Ripley, 2002). La technique de MDS[1] est implémentée dans le logiciel d'analyse statistique R[2]. Elle permet d'analyser une matrice pour un ensemble de données, disposées en rangées, à partir de leurs valeurs pour plusieurs variables, disposées en colonnes. L'objectif est de modéliser les similarités ou dissimilarités entre les données à partir de leurs valeurs afin de visualiser ces données dans un espace à deux dimensions. Cette technique vise à réarranger les données de façon efficace afin d'obtenir la configuration visuelle qui représente le mieux possible les distances observées entre les données. La meilleure représentation visuelle est celle qui maximise la qualité de l'ajustement (*goodness-of-fit*) et qui minimise la distorsion, en réduisant l'ensemble des dimensions aux deux dimensions visualisées (*plot*). La qualité de cette représentation visuelle est évaluée à l'aide du *stress*. Un pourcentage de stress plus faible indique une

[1] Le MDS est une méthode d'analyse multivariée descriptive, telle que l'analyse factorielle des correspondances (AFC) ou l'analyse en composantes principales (ACP). A la différence de ces techniques, le MDS permet d'analyser tout type de matrice de (dis)similarité, si les (dis)similarités sont évidentes. Le MDS n'impose pas de restrictions, telles que des relations linéaires entre les données sous-jacentes, leur distribution normale multivariée ou la matrice de corrélation (http://www.statsoft.com/textbook/stmulsca.html) [consulté le 04/04/2013].

[2] www.r-project.org [consulté le 04/04/2013].

© ATALA

meilleure représentation des distances observées entre les données. En règle générale, un pourcentage de stress inférieur à 10% est excellent et un pourcentage supérieur à 15% est inacceptable (Clarke, 1993 ; Borg et Groenen, 2005)[3]. Le logiciel R propose deux types de positionnement multidimensionnel[4], à savoir le `isoMDS`[5] et le `cmdscale`. Comme le isoMDS est un type de positionnement non métrique, il est plus flexible. Les deux types de positionnement prennent comme point de départ une matrice de dissimilarité pour calculer et visualiser la distance de chaque donnée par rapport aux autres. Cette matrice de dissimilarité est générée dans R, à partir de la matrice de base. Le résultat de l'analyse MDS est une visualisation (*plot*) selon deux axes, qui présente les données analysées et leurs distances, les regroupements de données et les données isolées ou périphériques.

L'analyse MDS permet de regrouper et visualiser en 2D la distance entre les cooccurrents de premier ordre, en fonction de leur degré d'association (valeur de LLR ou de PMI) avec les cooccurrents de deuxième ordre. Le point de départ est une matrice par mot-pôle dans laquelle les cooccurrents de premier ordre (*c*) constituent les rangées et les cooccurrents de deuxième ordre (*cc*) les colonnes. Il s'agit d'une matrice $c \times cc$ (lignes × colonnes), avec tous les *c* et *cc* pertinents. Les cellules de la matrice contiennent le degré d'association entre le *c* et le *cc* en question. La matrice de base ($c \times cc$) est réalisée à l'aide de scripts en Python, à partir de la base de données de cooccurrences de premier et deuxième ordre. Celle-ci a servi précédemment au calcul du degré de monosémie et rassemble toutes les informations requises, notamment le degré d'association. Notons que la matrice de base $c \times cc$ est creuse : de nombreux *cc* sont partagés par très peu de *c*.

2.2.2 Métriques pour la matrice de dissimilarité

À partir de la matrice de base, nous générons une matrice de dissimilarité, en calculant les distances par paire d'observations. Cette matrice est comparable à la matrice des distances kilométriques entre les villes, par exemple, pour les situer visuellement les unes par rapport aux autres. La métrique par défaut (`dist`) est la distance euclidienne, qui calcule le carré de la différence entre les valeurs de deux observations. La matrice de dissimilarité est ensuite utilisée pour le MDS et la visualisation des observations, en l'occurrence les cooccurrents de premier ordre en 2D.

Toutefois, la métrique utilisée pour réaliser la matrice de dissimilarité a une influence sur le regroupement et la visualisation. La distance euclidienne mesure la distance spatiale ou la distance en ligne droite entre les deux observations dans un espace 2D. La similarité entre les observations peut donc être exprimée en termes de distance. Les observations avec des valeurs similaires se caractérisent par une faible distance euclidienne. Lorsque les valeurs sont plus élevées, la distance sera plus grande. Les *c* avec des valeurs d'association très élevées avec un ou plusieurs *cc*, auront donc des valeurs de dissimilarité d'autant plus élevées en raison du carré de la différence. Dans la visualisation, ils se retrouveront à des distances plus éloignées ou même à des positions périphériques. De même, les *c* avec des valeurs d'association très faibles vont se retrouver ensemble, en dépit des différences entre leurs valeurs d'association avec les *cc*. Dans la matrice de dissimilarité, ces valeurs faibles

[3] http://geai.univ-brest.fr/~carpenti/2006-2007/Documents-R/MDS-non-metrique.html [consulté le 04/04/2013].
[4] Dans R : `library(MASS)`.
[5] Dans le logiciel R, la technique de isoMDS est appelée « Kruskal's Non-metric Multidimensional Scaling » : http://stat.ethz.ch/R-manual/R-patched/library/MASS/html/isoMDS.html [consulté le 04/04/2013].

© ATALA

perdront tout leur poids face aux valeurs plus élevées. Par conséquent, pour des matrices avec des valeurs d'association faibles et élevées, la métrique de la distance euclidienne ne convient pas très bien. Une métrique alternative pour générer la matrice de dissimilarité est l'angle du cosinus (*cosinus angle*). Celle-ci est couramment utilisée en sémantique distributionnelle (Padó et Lapata, 2007 ; Sahlgren, 2008). Elle s'applique à des observations représentées par des vecteurs et elle détermine la similarité en calculant l'angle entre les vecteurs. Des observations avec des vecteurs similaires se situent plus près dans l'espace multidimensionnel et se caractérisent donc par un angle plus petit. Même si l'ordre de grandeur des valeurs dans les vecteurs change, l'angle entre les vecteurs n'est pas affecté (van der Laan et Pollard, 2003). En analyse sémantique, il est souvent fait appel aux modèles vectoriels et à l'angle du cosinus pour établir la similarité entre deux documents ou entre deux mots (Peirsman et Geeraerts, 2009).

Nous pourrions donc considérer l'angle du cosinus[6] comme métrique alternative pour générer la matrice de dissimilarité. Les lignes (*c*) de la matrice de base sont conçues comme des vecteurs, comprenant une valeur par colonne. Pour ces vecteurs, la similarité (ou la dissimilarité) est calculée en fonction des valeurs d'association dans les différentes colonnes (*cc*). La fonction `distancematrix` permet de choisir la métrique[7], la distance euclidienne ou l'angle du cosinus, et de comparer les résultats visuels en fonction de la métrique choisie.

2.2.3 Problème de rareté des données

Comme nous l'avons déjà signalé ci-dessus, la matrice de base par mot-pôle souffre d'un problème de rareté de données. Par conséquent, la représentation visuelle est basée sur des informations très dispersées et de ce fait moins intéressantes. Il est possible d'enrichir la matrice $c \times cc$ par mot-pôle en faisant appel à des cooccurrences d'un ordre supérieur, en l'occurrence des cooccurrents de troisième ordre (ou *ccc*) (Bieman et al., 2004 ; Lemaire et Denhière, 2006). Les informations (sémantiques) apportées par les cooccurrences d'un ordre supérieur sont plus riches et plus robustes (Schütze, 1998). Il s'agit donc d'établir une matrice $c \times ccc$, avec tous les *c* pertinents disposés en rangées et tous les *ccc* pertinents (pour tous les *c* et tous les *cc*) en colonnes. Dans cette matrice $c \times ccc$, la valeur d'une case n'est pas simplement la valeur d'association entre le *c* et le *cc*, mais la somme de colonne d'une nouvelle matrice générée pour chaque *c* du mot-pôle, avec les *cc* en rangées et les *ccc* en colonnes. S'il y a *n ccc* au total pour tous les *cc* d'un *c*, cette nouvelle matrice $cc \times ccc$ permet de calculer la somme par colonne pour générer un vecteur à *n* dimensions. Ce vecteur à *n* dimensions permet de remplir les *n* cases de la rangée *c* de la matrice $c \times ccc$. Cette nouvelle matrice est moins creuse et donc plus intéressante pour visualiser les *c* en fonction de leurs informations sémantiques, soit tous les *ccc* de tous les *cc*.

2.2.4 Configuration préconisée ?

Il n'est pas évident de déterminer la meilleure configuration pour les analyses MDS de regroupement et de visualisation. Deux configurations sont envisageables pour la matrice de base par mot-pôle : la matrice $c \times cc$ et la matrice $c \times ccc$. Plusieurs métriques sont disponibles pour générer la matrice de dissimilarité, notamment la distance euclidienne et l'angle du cosinus. Par ailleurs, il existe plusieurs mesures d'association pour relever les

[6] Dans R : `library(hopach)`.
[7] Dans R : `d="euclid"` ou `d="cosangle"`.

© ATALA

cooccurrents pertinents, à savoir le LLR et la PMI (cf. section 2.1.1), parmi tant d'autres, comme le chi-carré et le t-score. La mesure du LLR s'accompagne d'une valeur p, qui permet de déterminer un seuil de significativité plus ou moins sévère. La mesure de la PMI n'est pas accompagnée d'une valeur p, mais le seuil pourrait être défini en fonction de la co-fréquence.

	Mesure d'association pour les cooccurrences	Métrique pour la matrice de dissimilarité	Seuil pour le repérage / le maintien des cooccurrences
Matrice $c \times cc$	LLR	Distance euclidienne	Valeur $p \leq 0{,}0001$
		Angle du cosinus	Valeur $p \leq 0{,}0001$
	PMI	Distance euclidienne	Plusieurs seuils inférieurs et supérieurs
		Angle du cosinus	Plusieurs seuils inférieurs et supérieurs
Matrice $c \times ccc$	PMI	Distance euclidienne	Plusieurs seuils inférieurs et supérieurs
		Angle du cosinus	Plusieurs seuils inférieurs et supérieurs

TABLE 1 – Aperçu des différentes configurations pour les analyses MDS

Dans nos recherches actuelles, nous essayons de déterminer la configuration la plus intéressante pour l'interprétation sémantique de la dispersion des cooccurrents de premier ordre (*c*) et donc pour l'interprétation sémantique du mot-pôle (cf. table 1).

Pour la configuration avec la mesure d'association du LLR, nous gardons le seuil de significativité sévère (valeur $p \leq 0{,}0001$) de notre analyse sémantique quantitative afin de pouvoir comparer les résultats des analyses MDS. Ce seuil de significativité a comme avantage d'exclure, au moment du repérage des cooccurrents significatifs, les mots grammaticaux très fréquents mais peu pertinents (articles, pronoms, etc.).

Pour la mesure d'association de la PMI, nous prenons en considération plusieurs seuils inférieurs et supérieurs. La prise en compte du seuil inférieur permet de relever uniquement des cooccurrences pertinentes et d'exclure des combinaisons de mots trop peu fréquentes. Ainsi, le seuil inférieur permet de remédier au problème de la mesure de PMI, qui surestime la cooccurrence de mots rares. Pour le seuil inférieur, nous considérons trois valeurs de co-fréquence minimale, à savoir 5, 10 et 20. Le seuil supérieur vise à supprimer les mots grammaticaux parmi les cooccurrents de premier ordre visualisés sur la représentation visuelle (*plot*) de l'analyse MDS (cf. section 3). Parmi les cooccurrents de deuxième et troisième ordre, les mots grammaticaux sont préservés, parce qu'ils contribuent à la

 © ATALA

différenciation sémantique. Pour le seuil supérieur, nous envisageons plusieurs critères, afin d'aboutir à la visualisation la plus claire et la plus efficace :

1. pas de seuil supérieur,
2. co-fréquence des cooccurrents de premier ordre avec le mot-pôle,
3. fréquence absolue des cooccurrents de premier ordre,
4. suppression manuelle des mots grammaticaux dans les matrices $c \times cc$ et $c \times ccc$.

Les matrices $c \times cc$ et $c \times ccc$ dans les différentes configurations sont soumises à une analyse MDS dans le logiciel R. Un pourcentage de stress inférieur à 15% permet de retrouver les configurations significatives et intéressantes.

3 Discussion des résultats

3.1 Matrice $c \times cc$

Les premières expérimentations sont conduites pour les cooccurrences du mot-pôle *tour*. Pour la configuration de la matrice $c \times cc$, avec le LLR et la distance euclidienne, le stress est très faible (0,047%), mais la répartition visuelle des *c* en fonction des *cc* n'est pas utile. Les 3 *c* périphériques se caractérisent par des fréquences et co-fréquences très élevées et par des valeurs d'association (LLR) gonflées. Les autres *c* moins fréquents se regroupent et ne permettent pas de pointer vers des sens différents. Les 3 *c* périphériques constituent une unité polylexicale avec le mot-pôle, à savoir *tour à commande numérique*. Ces 3 *c* sont périphériques, mais dans une direction opposée, en raison de leur cooccurrence pertinente avec des *cc* différents. Il est clair que la valeur d'association et la métrique de la distance euclidienne ne sont pas appropriées. La métrique de l'angle du cosinus réussit à mieux répartir les *c*, mais cette analyse MDS n'est pas acceptable à cause du stress élevé (20,35%).

La mesure d'association de la PMI permet de remédier au problème des fréquences élevées. La prise en compte d'un seuil inférieur présente l'avantage d'éviter des problèmes de surestimation pour les mots rares. Pour un seuil de co-fréquence supérieur à 10 et sans seuil supérieur, le MDS avec la distance euclidienne est acceptable (stress de 9,27%). Toutefois, le résultat visuel reste difficile à interpréter en raison des nombreux mots grammaticaux parmi les cooccurrents de premier ordre (cf. figure 1). Les mots lexicaux se regroupent pour constituer un grand cluster lexical, à l'exception des mots *outils* et *broche*.

Afin de mieux répartir les *c* et d'éviter le regroupement des *c* moins fréquents, nous recourons de nouveau à l'angle du cosinus, mais le stress est supérieur à 15% pour les trois seuils inférieurs (co-fréquence $\geq$ 5, 10 et 20). Il s'avère que la distorsion des distances entre les observations pour la visualisation en 2D est trop importante. Plusieurs critères sont envisageables. La co-fréquence des *c* avec le mot-pôle (p.ex. co-fréquence $\geq$ 500 ou 250) permet de supprimer beaucoup de mots grammaticaux dans la matrice $c \times cc$, mais aussi les mots lexicaux les plus importants, comme *broche, inversée, verticaux, CNC*. Nous optons dès lors pour une « coupe intelligente », c'est-à-dire un *cut off* à 500 ou 250, mais avec maintien des mots lexicaux. Cela revient à enlever manuellement les mots grammaticaux, ayant une co-fréquence élevée. Cependant, aussi parmi les co-fréquences moins élevées, on recense bon nombre de mots grammaticaux qui rendent la visualisation des *c* trop compliquée. Il est plus judicieux d'enlever manuellement tous les mots grammaticaux dans la matrice $c \times cc$. Il en va de même pour le critère de la fréquence absolue des *c*.

 © ATALA

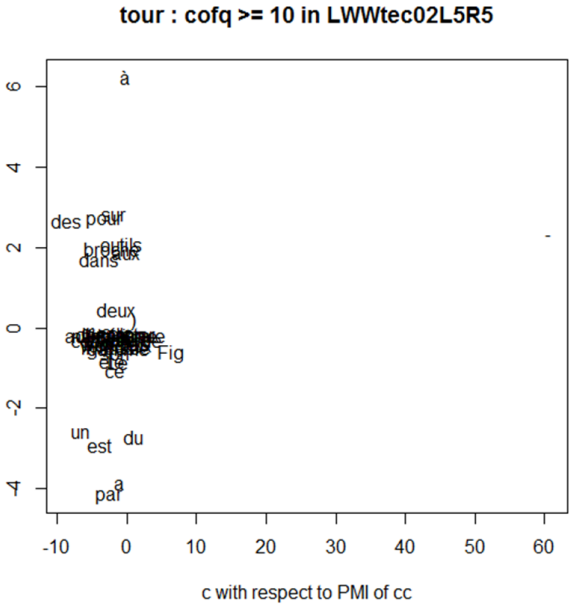


FIGURE 1 – MDS des *c* de *tour* par rapport aux *cc* (PMI – distance euclidienne).

3.2 Matrice enrichie *c* × *ccc*

Dans la plupart des configurations, c'est le stress qui pose le problème le plus important. Pour la distance euclidienne, il frôle les 15%, mais pour l'angle du cosinus il oscille presque partout entre 28% et 36% (pour la PMI). Cela s'explique probablement par les propriétés de la matrice considérée (*c* × *cc*), qui souffre d'un véritable problème de rareté de données. Pour enrichir cette matrice, nous recourons aux *ccc*. Les configurations envisagées dans nos expérimentations pour cette matrice enrichie *c* × *ccc* font intervenir uniquement la mesure d'association de la PMI. Elles s'appuient sur les deux métriques pour générer la matrice de dissimilarité. Elles font varier le seuil inférieur (5, 10, 20) et le seuil supérieur, tout comme les configurations pour la matrice *c* × *cc*. Pour deux configurations (*c* × *ccc*), le stress est acceptable. Il s'agit des configurations avec un seuil inférieur de 5 et 10 (co-fréquence ≥ 5 et 10), après suppression manuelle des mots grammaticaux. Le stress y est respectivement de 13,25% et de 14,81%. Notons que les numéraux cardinaux ne sont pas supprimés pour le mot-pôle *tour*. En effet, ils entrent dans un certain nombre d'unités polylexicales pertinentes, telles que *tour à trois axes, tour à trois broches*. Dans la matrice *c* × *ccc*, seules les lignes avec des *c* lexicaux sont soumises à l'analyse MDS. La visualisation ci-dessous (cf. figure 2) qui montre la répartition des *c* à partir de leur valeur d'association (PMI) avec les *ccc* permet de formuler les premières conclusions sémantiques intéressantes. Il n'y a plus de cluster lexical et les *c* sont répartis dans l'espace 2D.

Le *c* le plus périphérique, *horizon*, est une indication d'un des sens généraux du mot-pôle *tour* (p.ex. *un tour d'horizon*). Le *c mille* occupe également une position plus périphérique et renvoie au sens « rotation » (p.ex. *dix mille tours par minute*). *CNC*, *commande* et *numérique*

© ATALA

se retrouvent ensemble : ils sont sémantiquement liés. La figure 3 ci-dessous visualise l'analyse MDS pour un seuil inférieur de co-fréquence ≥ 5 et recense plus de *c*. La répartition et le regroupement de ces *c* permettent de préciser les conclusions formulées pour la visualisation précédente (cf. figure 2). Pour des raisons de fiabilité statistique de la mesure de la PMI, un seuil inférieur d'une co-fréquence inférieure à 5 n'est pas envisageable.

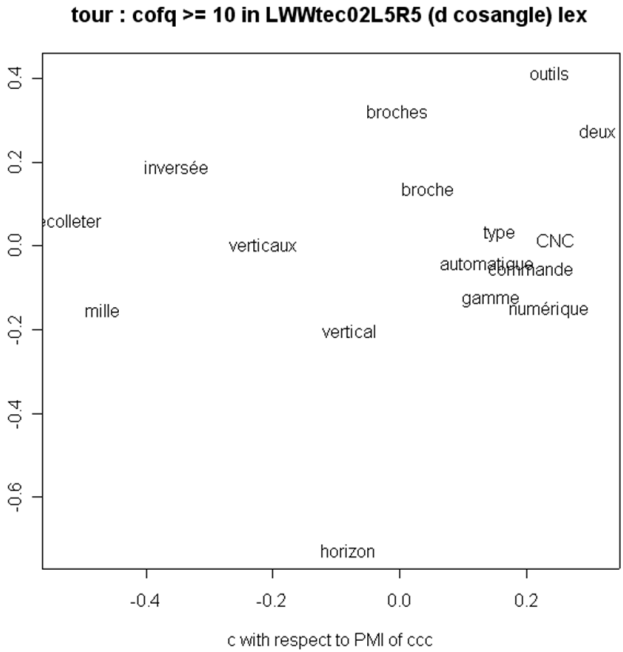


FIGURE 2 – MDS des *c* de *tour* par rapport aux *ccc* (PMI – angle du cosinus) : mots lexicaux au seuil inférieur de co-fq ≥ 10.

Dans la figure 3 ci-dessous, la répartition des *c* représente bien le caractère sémantiquement hétérogène de *tour*. Les proximités et les distances sémantiques correspondent à notre objectif de recherche, qui consiste à mieux comprendre l'hétérogénéité sémantique calculée par la mesure de recoupement. Nous observons des *c* périphériques, qui pointent vers des sens particuliers, tels que *horizon* (sens général dans « tour d'horizon ») et *mille* (sens technique dans « mille tours par minute »). Dans la partie supérieure à droite, on retrouve des *c* qui attestent le sens technique particulier « tour inversé », avec *inversés, inversé* et *bibroches*. Le nuage de *c* dans la partie inférieure à gauche visualise le sens technique « machine-outil pour l'usinage des pièces ». La présence de plusieurs petits regroupements de *c* témoigne de la variation des contextes d'emplois. On observe un contexte d'emploi spécialisé « tour (à) deux broches » à gauche en bas : *axes, broches, broche, outils, deux, trois, centres* et même au milieu *vertical* et *verticaux*. À gauche au milieu on retrouve *commande, numérique, CNC, gamme, série, type* et quelques *c* moins spécialisés. Les *c* plus généraux comme *manutention, ligne, générale, production, conception* se retrouvent dans la partie supérieure de ce nuage et font preuve d'un contexte d'emploi plus général. Cette visualisation interprétable est réalisée grâce à la matrice *c* × *ccc* enrichie.

© ATALA

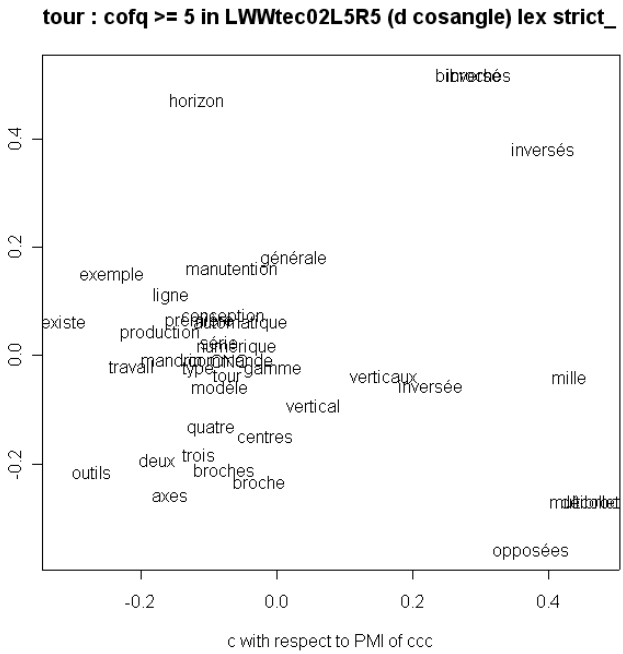


FIGURE 3 – MDS des *c* de *tour* par rapport aux *ccc* (PMI – angle du cosinus) : mots lexicaux au seuil inférieur de co-fq ≥ 5.

4 Conclusion et recherches futures

Ces premières expérimentations pour l'exploration visuelle de la proximité sémantique entre les cooccurrents du mot-pôle *tour* nous ont permis de formuler des conclusions intéressantes sur son hétérogénéité sémantique. La répartition des cooccurrents de premier ordre avec quelques regroupements et cooccurrents isolés montre bien les différents sens de ce mot à la fois homonymique et polysémique. Nous avons étudié l'effet de plusieurs configurations de mesures d'association, de métriques de distance, de seuils inférieurs et supérieurs dans le but de trouver la meilleure configuration. Nous aimerions ainsi mettre au point une méthodologie opérationnelle et interprétable pour visualiser les *c* d'un mot-pôle en fonction des informations de cooccurrence avec les *c* et/ou les *ccc* et pour comprendre ainsi ce qui se passe derrière notre mesure de monosémie.

Dans nos recherches futures, nous envisageons d'examiner plusieurs mots-pôles afin de peaufiner la méthodologie développée jusqu'ici et de parvenir à la meilleure configuration de paramètres. Nous visons aussi à travailler sur l'ensemble du corpus technique de 1,7 million d'occurrences. Finalement, nous devrons nous atteler aux unités polylexicales, en faisant intervenir des critères syntaxiques pour le repérage des unités polylexicales spécifiques du corpus technique. À l'instar de notre étude sémantique quantitative, cette étude de regroupement et de visualisation se limite pour l'instant aux unités lexicales simples.

© ATALA

Références

BERTELS, A. et SPEELMAN, D. (2012). La contribution des cooccurrences de deuxième ordre à l'analyse sémantique. *In Corpus* n°11, pages 147-165.

BERTELS, A., SPEELMAN, D. et GEERAERTS, D. (2010). La corrélation entre la spécificité et la sémantique dans un corpus spécialisé. *In Revue de Sémantique et de Pragmatique* n°27, pages 79–102.

BIEMANN, C., BORDAG, S. et QUASTHOFF, U. (2004). Automatic acquisition of paradigmatic relations using iterated co-occurrences. *In Proceedings of LREC 2004*. Lisboa, Portugal, pages 967–970.

BLUMENTHAL, P. et HAUSMANN, F.J., éditeurs (2006). *Collocations, corpus, dictionnaires. Langue française*, n° 150.

BORG, I. et GROENEN, P. (2005). *Modern Multidimensional Scaling: theory and applications* (Second edition). New York. Springer-Verlag.

CHURCH, K.W. et HANKS, P. (1990). Word association norms, mutual information, and lexicography. *In Computational Linguistics* n°16(1), pages 22-29.

CLARKE, K.R. (1993). Non-parametric multivariate analyses of change in community structure. *In Australian Journal of Ecology* n°18, pages 117-143.

CONDAMINES, A., éditeur (2005). *Sémantique et corpus*, Paris. Hermès-Science.

COX, T.F. et COX, M.A.A. (2001). *Multidimensional Scaling*. Boca Raton, FL. Chapman & Hall.

DUNNING, T. (1993). Accurate methods for the statistics of surprise and coincidence. *In Computational Linguistics* n°19(1), pages 61–74.

ERIKSEN, L. 2002. Die Polysemie in der Allgemeinsprache und in der juristischen Fachsprache. *In Hermes – Journal of Linguistics* n°28, pages 211–222.

EVERT, S. (2007). *Corpora and collocations.* Extended Manuscript of Chapter 58 of Lüdeling A. et Kytö M., 2008, *Corpus Linguistics. An International Handbook,* Mouton de Gruyter, Berlin.

FABRE, C., HABERT, B. et LABBÉ, D. (1997). La polysémie dans la langue générale et les discours spécialisés, *In Sémiotiques* n°13, pages 15-30.

FERRARI, L. (2002). Un caso de polisemia en el discurso jurídico? *In Terminology* n°8(2), pages 221–244.

FERRET, O. (2010) Similarité sémantique et extraction de synonymes à partir de corpus. *In Proceedings of TALN 2010*, Montréal, Canada.

GREFENSTETTE, G. (1994). Corpus-derived first, second and third-order word affinities. *In Proceedings of Euralex '94.* Amsterdam, pages 279-290.

GROSSMANN, F. et TUTIN, A., éditeurs (2003). *Les collocations, analyse et traitement, Travaux et Recherches en linguistique appliquée*, Série E, vol. 1.

HABERT, B., ILLOUZ, G., FOLCH, H. (2004). Dégrouper les sens : pourquoi ? comment ? *In Actes des JADT 2004,* Louvain-la-Neuve, pages 565–576.

 © ATALA

HABERT, B., ILLOUZ, G., FOLCH, H. (2005). Des décalages de distribution aux divergences d'acception, *In* (Condamines, 2005), pages 277–318.

HARRIS, Z. S. (1991). *A theory of language and information. A mathematical approach.* Oxford. Oxford University Press.

KRUSKAL, J.B. et WISH, M. (1978). *Multidimensional Scaling*. Sage University Paper series on Quantitative Applications in the Social Sciences, number 07-011. Newbury Park, CA. Sage Publications.

LEMAIRE, B. et DENHIÈRE, G. (2006). Effects of High-Order Co-occurrences on Word Semantic Similarity. *Current Psychology Letters* n°18(1). `http://cpl.revues.org/index471.html`. [consulté le 04/04/2013].

MARTINEZ, W. (2005), *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*, Thèse de Doctorat en Sciences du Langage, Université de la Sorbonne nouvelle - Paris 3, Paris.

MAYAFFRE, D. (2008), Quand 'travail', 'famille', 'patrie' co-occurrent dans le discours de Nicolas Sarkozy. Etude de cas et réflexion théorique sur la co-occurrence, *In Actes des JADT 2000,* Lyon, pages 811-822.

PADÓ, S. et LAPATA, M. (2007). Dependency-based construction of semantic space models. *In Computational Linguistics* n°33(2), pages 161–199.

PEIRSMAN, Y. et GEERAERTS, D. (2009). Predicting Strong Associations on the Basis of Corpus Data. *In Proceedings of EACL-2009*, Athens, Greece, pages 648-656.

ROY, T. (2007) *Visualisations interactives pour l'aide personnalisée à l'interprétation d'ensembles documentaires*, Thèse de doctorat de l'Université de Caen Basse-Normandie, France.

SAHLGREN, M. (2006). *The Word-Space Model*. Ph.D. thesis, Stockholm University, Sweden.

SAHLGREN, M. (2008). The Distributional Hypothesis. *In Rivista di Linguistica* n°20(1), pages 33–53.

SCHÜTZE, H. (1998), Automatic Word Sense Discrimination. *In Computational Linguistics* n°24(1), pages 97-123.

TEMMERMAN, R. (2000). *Towards new ways of terminology description. The sociocognitive approach,* Amsterdam/Philadelphia. John Benjamins Publishing Company.

TURNEY, P.D. et PANTEL, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *In Journal of Artificial Intelligence Research* n°37, pages 141-188.

VAN DER LAAN, M.J. et POLLARD, K.S. (2003). A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. *In Journal of Statistical Planning and Inference* n°117, pages 275-303.

VENABLES, W.N. et RIPLEY, B.D. (2002). *Modern Applied Statistics with S*, (Fourth edition). New York. Springer-Verlag.

WÜSTER, E. (1991). *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*, (3. Aufl.), Bonn. Romanistischer Verlag.

 © ATALA

Étude des relations sémantiques dans les reformulations de requêtes sous la loupe de l'analyse distributionnelle

Clémentine Adam Cécile Fabre Ludovic Tanguy
CLLE-ERSS : CNRS & Université de Toulouse
5, Allées Machado, 31058 Toulouse Cedex 9
adam@univ-tlse2.fr, cfabre@univ-tlse2.fr, tanguy@univ-tlse2.fr

RÉSUMÉ

Dans cet article, nous confrontons une base distributionnelle construite à partir d'un corpus d'articles de revues de sciences humaines à des substitutions observées dans les journaux de requêtes du moteur interrogeant ce même corpus ; le recouvrement entre les deux types de données est important (59%). Ces résultats contribuent à deux pistes de recherche : d'une part nous montrons l'adéquation de la sémantique distributionnelle pour appréhender une large palette de relations sémantiques en jeu dans les reformulations de requêtes ; d'autre part, nous introduisons des données pouvant être exploitées pour l'évaluation de ressources distributionnelles de manière bien plus satisfaisante que par la comparaison avec des *gold standards* tels que des dictionnaires de synonymes.

ABSTRACT

Studying semantic relations in query reformulation under the scope of distributional analysis

In this paper, we compare a distributional resource built from a corpus of humanities and social sciences academic papers to substitutions recorded in user query logs covering the same corpus. We observed a good overlap between the two datasets (59%). These results show that distributional semantics is a fitting tool to analyze the wide variety of semantic relations involved in query reformulation. Moreover, the method that we introduce may be used for distributional resources evaluation, and is better fitted to this task than comparison with gold standards.

MOTS-CLÉS : sémantique distributionnelle, Recherche d'information, reformulation de requêtes.

KEYWORDS: distributional semantics, information retrieval, query reformulation.

1 Introduction

Les ressources lexicales construites par des méthodes distributionnelles ont été confrontées de différentes manières à l'expertise humaine pour chercher à cerner la nature des relations qui sont détectées par ce biais. La comparaison à d'autres ressources lexicales servant de *gold standard* – dictionnaires de synonymes et thésaurus – (Weeds, 2003; Anguiano *et al.*, 2011) a montré que les liens de proximité sémantique contruits de manière distributionnelle ne recouvrent que très partiellement les relations décrites dans les ressources lexicales existantes. Ainsi, (Morlane-Hondère et Fabre, 2012) montrent qu'on retrouve dans les couples de voisins distributionnels calculés à partir d'un corpus issu de Wikipédia 30 à 40% des couples contenus

 © ATALA

dans un dictionnaire de synonymes, mais que ces synonymes ne constituent qu'une infime proportion de l'ensemble des liens de voisinage distributionnel. Ces proportions sont comparables lorsque la confrontation porte sur le réseau lexical français Jeux de mots (Lafourcade et Joubert, 2010), qu'il s'agisse des relations d'antonymie, d'hyperonymie, ou de méronymie. Par conséquent, l'utilisation de telles ressources n'éclaire qu'une toute petite partie seulement des relations sémantiques calculées. Deux raisons expliquent cette intersection limitée : le fait que d'autres relations de sens plus lâches, souvent qualifiées de "non classiques" (Morris et Hirst, 2004) et qui n'entrent ordinairement pas dans le périmètre des ressources lexicales existantes, sont détectées par la proximité distributionnelle (cohyponymie, mots de même domaine, etc.) ; le fait également que les relations acquises à partir de corpus peuvent être fortement contextualisées et par conséquent ne pas être recensées dans des lexiques. Une seconde approche consiste à confronter les résultats du calcul distributionnel au jugement humain à travers différentes tâches impliquant un jugement de similarité sémantique : détection de synonymes, *priming*, catégorisation, analogie (Padó et Lapata, 2007; Baroni et Lenci, 2010). Ces expériences montrent que la proximité distributionnelle offre une bonne approximation du jugement de similarité, et qu'une palette importante de relations sont détectées par la méthode distributionnelle.

Dans le travail présenté ici, notre objectif est de confronter une ressource distributionnelle à des données obtenues dans le contexte d'une tâche de recherche d'information. L'utilisation de ressources distributionnelles en RI a donné lieu à différentes expériences visant particulièrement l'expansion automatique de requêtes, depuis (Grefenstette, 1994) et, plus récemment, (van der Plas et Tiedemann, 2008; Picton *et al.*, 2008). Ces deux études, qui exploitent les données des campagnes d'évaluation CLEF, font état de résultats globalement décevants, avec des augmentations de performance non significatives. Elles montrent également par des analyses plus locales que les effets de l'expansion sont très contrastés, selon le type de requête, la catégorie de mots expansés et la nature du lien sémantique : les deux études mettent au jour des faiblesses spécifiques concernant l'expansion des adjectifs et des noms propres, liées en particulier à la génération de liens de cohyponymie qui introduisent du bruit (ex : lien de *russe* vers *allemand* ou *chinois*). L'exploitation de liens sémantiques non classiques (ex : *grève*/*réclamer*, *clonage*/*éthique*), et plus spécifiquement de liens entre des mots de même domaine (comme *bombe*, *guerre*) améliore par contre nettement le traitement de certaines requêtes. Une utilisation en aveugle de ressources distributionnelle s'avère donc inefficace. Une meilleure connaissance des types de requêtes et des relations de sens mises en jeu dans ce type de tâche s'avère nécessaire pour permettre une utilisation plus adaptée d'informations sémantiques en RI.

L'étude que nous présentons ici se situe en amont de ce type de travaux et se fonde sur des données écologiques : il s'agit de journaux de requêtes produites par les utilisateurs d'un moteur de recherche. L'objet de l'article est d'étudier le recouvrement entre les liens sémantiques exploités dans les reformulations, et ceux qui sont calculés sur la base du critère de proximité distributionnelle. Nous présentons d'abord les caractéristiques générales de ces données (section 2), avant de présenter l'échantillon de requêtes sur lequel se base notre étude : issu du moteur de recherche d'`openedition.org`, il est composé de couples de requêtes correspondant à des reformulations fondées sur la substitution d'un mot (section 3.1.). Nous présentons ensuite la base distributionnelle que nous avons construite à partir de la base de textes interrogée par les utilisateurs (section 3.2.). Dans la dernière section, nous présentons les résultats de cette comparaison, qui montrent un recouvrement important entre les deux sources de données (mots substitués / voisins distributionnels), et nous analysons certaines sources de décalage permettant de mieux comprendre les différences entre ces deux types de proximité sémantique.

 © ATALA

2 Les reformulations de requêtes

L'étude des journaux des requêtes soumises à un moteur de recherche (ou *query logs*) a, dès le début, suscité un grand intérêt pour la communauté de la recherche d'information, que ce soit pour mieux comprendre le comportement des utilisateurs ou pour modéliser celui-ci afin d'améliorer la réponse d'un moteur de recherche (Jansen *et al.*, 2000).

Parmi les différents phénomènes abordés dans ces études, la notion de *reformulation* est l'une des plus riches en information sur la nature des interactions entre un utilisateur et un système de recherche (Jansen *et al.*, 2009). Une reformulation est simplement la soumission au moteur de recherche d'une seconde requête, différente de la précédente, mais qui cherche à remplir le même besoin d'information. Le temps passé et les actions de l'utilisateur entre les deux requêtes peuvent varier, mais on a pu dégager empiriquement les principales modifications formelles que subissent les textes des deux requêtes. En nous inspirant de l'état de l'art établi par (Huang et Efthimiadis, 2009), nous les présentons dans la table 1 avec des exemples extraits de notre propre collection de données.

Modification	Exemples
Correction orthographique	`praglatismes` → `pragmatisme`
Modification de l'espacement	`église catholiquefinancement` → `église catholique financement`
Réordonnancement	`monde rural france survivance` → `survivance monde rural france`
Ajout de mot(s)	`algerie` → `algerie femmes`
Suppression de mot(s)	`hyperactivité à l'ecole` → `hyperactivité`
Remplacement de mot(s)	`stéréotypes sexistes` → `stéréotypes machistes`

TABLE 1 – Différents types de schémas de reformulation

D'un point de vue plus profond, les reformulations qui modifient de façon non triviale l'expression de la requête peuvent être classées suivant les 4 types suivants, qui pour certains se traduisent par plusieurs mécanismes de surface : *généralisation*, *spécification*, *reformulation* et *mouvement parallèle*. Des exemples de telles opérations sont présentés dans la table 2. La généralisation correspond clairement à un élargissement du champ de recherche visant à réduire le silence, et la spécification à son rétrécissement pour réduire le bruit. Ces deux opérations peuvent se traduire chacune par deux opérations formelles (remplacement ou ajout/suppression de mots). La reformulation correspond à une simple paraphrase, alors que le mouvement parallèle est au contraire une modification importante de la requête, en considérant une alternative [1].

Comme on peut le voir dans ces quelques exemples, ces reformulations sont a priori le lieu d'expression de relations sémantiques très variées entre les requêtes (ou entre les termes de celles-ci). Plus particulièrement, on peut voir dans la table 2 que les opérations de substitution d'un mot par un autre peuvent se faire en suivant (au moins) quatre relations sémantiques : hyponymie, hyperonymie, synonymie et co-hyponymie. Dans le cas de cette substitution, le fait qu'une partie de la requête reste inchangée est un indicateur fiable de la continuité sémantique entre les deux requêtes, même si la cible de la requête a pu être sensiblement modifiée.

Les requêtes sur lesquelles nous basons cette étude, et dont sont extraits les exemples présen-

1. Les exemples canoniques du mouvement parallèle sont, dans le cas des requêtes soumises aux moteurs Web généralistes, le remplacement d'une marque ou d'un type de produit par un autre lors d'une recherche à visée transactionnelle, comme dans `nikon camera` → `canon camera` ou `billet train paris-toulouse` → `billet avion paris-toulouse`.

 © ATALA

Généralisation	Suppression de mots Remplacement par un hypéronyme	`inégalités sociales de santé` → `inégalités de santé` `sociologie arts martiaux` → `sociologie du sport`
Spécification	Ajout de mots Remplacement par un hyponyme	`faillite` → `théories de la faillite` `activités dans la montagne` → `escalades dans la montagne`
Reformulation	Remplacement par un synonyme	`territoire voiture` → `territoire automobile`
Mouvement parallèle	Remplacement par un co-hyponyme	`siège romain` → `siège gaulois`

TABLE 2 – Quatre types d'opération effectuée lors de la reformulation de requêtes

tés ci-dessus, sont celles qui ont été soumises au moteur de recherche du site OpenEdition (`www.openedition.org`) qui regroupe un grand nombre de ressources documentaires dans le domaine des sciences humaines et sociales. Nous disposons, grâce à la collaboration du CLEO (Centre pour L'édition Electronique Ouverte) de l'ensemble des requêtes soumises au moteur, ainsi que des adresses anonymisées qui nous permettent de suivre le comportement d'un utilisateur au cours d'une session (documents consultés, temps passé, etc.), voir (Leva, 2013) pour plus de détails.

3 Données et méthodologie

Dans cette section, nous décrivons les données que nous avons extraites et confrontées dans le cadre de cette étude : un ensemble de paires de requêtes présentant des remplacements lexicaux, extraites des journaux de requêtes d'OpenEdition (sous-section 3.1) et une base distributionnelle construite à partir de la collection de documents interrogée par le moteur de recherche d'OpenEdition (sous-section 3.2). Nous présentons ensuite la méthodologie de croisement de ces données (sous-section 3.3).

3.1 Les substitutions dans les journaux de requêtes

Pour l'extraction de paires de requêtes comprenant une substitution, nous avons utilisé un extrait des logs de recherche d'OpenEdition comprenant 194 363 requêtes réparties en 57 813 sessions de recherche de plus de deux requêtes (une session est ici définie grossièrement comme une séquence de requêtes soumises par le même utilisateur dans une même journée).

À partir de ces données, nous avons extrait toutes les paires de requêtes R_a, R_b telles que :
– R_a et R_b appartiennent à la même session et sont consécutives ;
– R_a et R_b comprennent un même nombre n de mots avec $n \geq 2$;
– R_a et R_b comprennent $n-1$ mots communs.
Nous nous basons donc sur une vision limitée de la substitution, qui ne porte ici que sur un seul mot dans des requêtes de plusieurs mots. Ces critères ont permis d'extraire un ensemble de 11321 paires de requêtes, ce qui représente plus de 8% des paires de requêtes du log analysé. Cela montre que le type de reformulations sur lequel nous avons choisi de travailler, bien que très circonscrit, n'est pas marginal dans les stratégies de recherche des utilisateurs.

 © ATALA

Nous fournissons dans la suite de cette sous-section quelques éléments de description des paires de requêtes extraites, dont des exemples sont répertoriés dans le tableau 3.

a	relations entre espace pouvoir et **société** relations entre espace pouvoir et **identité**
b	la gouvernance **internationalle** la gouvernance **international**
c	québec intégration **immigration** travail intégration **immigrants** travail québec
d	**securité** des sites de rencontres **dangers** des sites de rencontres
e	traduction **inginierie** traduction **technique**
f	**performance** système d'information **évaluer** système d'information
g	anthropologie des **iécoles** anthropologie des **institutions**
h	représentation **culturelle** de la douleur représentation **sociale** de la douleur
i	desavantages de l apprentissage d une **deuxieme** langue desavantages de l apprentissage d une langue **etrangere**
j	lieux de pouvoir paris élysée **louvre** lieux de pouvoir paris élysée **matignon**

k	cheminées **hydroterhmales** cheminées **hydrothermales**
l	societé **et** transition societé **en** transition
m	ville **de** paris **imaginaire** ville paris
n	**encourir** une condamnation **prononcer** une condamnation
o	**emploi** rural **travail** rural
p	climat **mousson** climat **tropical**
q	rire **thérapeutique** rire **guérir**
r	musique et **apprentissage** **musique** et **mémorisation** **rythmique** et mémorisation
s	stéréotypes **genrés** stéréotypes **sexistes** stéréotypes **machistes**

TABLE 3 – Exemples de paires de requêtes extraites

Positions des mots substitués Les paires de requêtes extraites comprennent entre 2 et 20 mots (moyenne : 2,95 mots). Lors de l'extraction, aucune incidence n'a été donnée à l'ordre des mots dans chaque requête. Il apparaît toutefois que celui-ci est généralement préservé entre les deux requêtes (peu de réordonnancements). La figure 1 résume les schémas de substitution observés : dans 92% des cas, le mot supprimé et le mot ajouté se trouvent à la même position, majoritairement à la fin de la requête (52%). Les changements de position (*cf.* tab. 3 ex. c et i) sont rares sans être marginaux : 8% des paires de requêtes sont concernées.

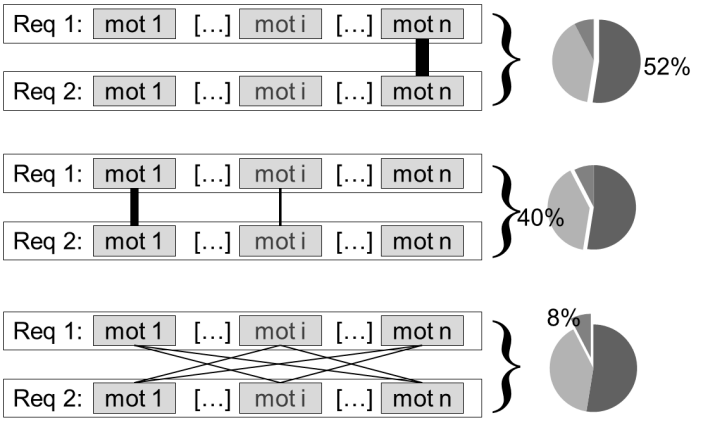


FIGURE 1 – Positions des mots substitués

© ATALA

Mots vides, corrections orthographiques et fautes non corrigées Certaines substitutions portent sur des mots vides (prépositions, déterminants, ...) : substitués entre eux (tab. 3 ex. l) ou avec un mot plein (ex. m) – dans ce dernier cas, il s'agit alors plutôt d'un ajout que d'un remplacement (le mot *imaginaire* est ajouté à la requête plus que substitué au mot *de*, qui n'avait pas d'impact sur la recherche). Nous avons repéré ces mots-vides à l'aide d'une *stop-list* ; 563 paires de requêtes sont concernées (moins de 5% de l'ensemble). Les corrections orthographiques, qui ne nous intéressent pas dans le cadre de cette étude, représentent quant à elles une très large partie des données extraites. Afin d'évaluer leur importance, nous avons appliqué une distance d'édition (ou distance de Levenstein) sur les paires de mots substitués, après exclusion des mots vides (les paires telles que `et` → `en` ne sont donc pas considérées comme des corrections orthographiques). Avec cette méthode, nous avons repéré 5155 corrections orthographiques (46% des données). Mais toutes les fautes d'orthographe ne donnent pas lieu à une correction, comme on peut le voir avec les substitutions `inginierie` → `technique` (ex. e) et `iécole` → `institution` (ex. g) ; ces paires exhibent un lien sémantique, mais sont de nature à faire buter les traitements. Pour repérer les mots incorrects, nous avons fait appel à deux lexiques de formes fléchies : Morphalou [2] et GLÀFF (Sajous *et al.*, 2013), ainsi qu'à une liste de noms propres extraite de Wikipédia. 1727 reformulations (15% des données) impliquent des mots qui ne sont recensés par aucune de ces trois ressources.

Catégories morpho-syntaxiques concernées À partir des données filtrées et lemmatisées par l'appel aux lexiques, nous avons étudié la représentation des catégories morpho-syntaxiques dans les reformulations de requêtes. La figure 2 résume les proportions observées. Les noms dominent largement, suivis des adjectifs, des noms propres et des verbes ; un très petit nombre d'adverbes a également été relevé (ex. `parler franchement` → `parler franc`). Les remplacements au sein d'une même catégorie, par exemple Nom→Nom (tab. 3 ex. a, c, d, etc.) sont logiquement majoritaires, Adjectif→Adjectif (ex. h, i). Toutefois, les substitutions intercatégorielles sont également très fréquentes : le tableau 3 donne à voir des exemples de substitutions Nom→Adj (ex. e et p), Nom→Verbe (ex. f) ou encore Adjectif→Verbe (ex. q).

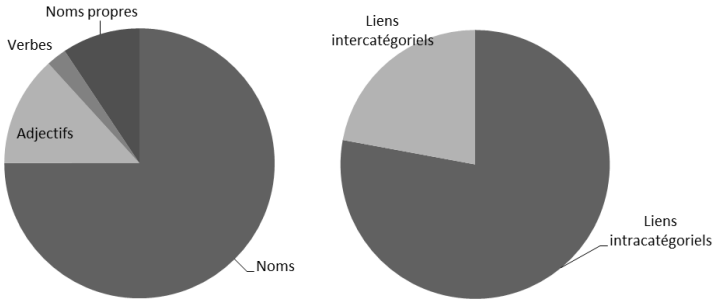


FIGURE 2 – Catégories morpho-syntaxiques des mots substitués

Variété des liens sémantiques Les reformulations qui nous intéressent plus particulièrement dans le cadre de cette étude sont celles qui présentent un lien sémantique entre les deux mots

2. développé par le CNRTL (ATILF) : `http://www.cnrtl.fr/lexiques/morphalou/`

© ATALA

substitués. Il s'agit donc, *a priori*, de toutes celles qui ne sont pas des corrections orthographiques ou des remplacements de mots vides, puisque nous faisons l'hypothèse que la partie inchangée lors de la reformulation est garante de la continuité sémantique entre les deux requêtes (*cf.* section 2). Les données présentées dans le tableau 3 donnent à voir une grande variété de relations sémantiques, allant au delà de ce qui est classiquement envisagé dans l'étude des reformulations de requêtes (hypo/hypéronymie, co-hyponymie, synonymie, *cf.* tableau 2). Outre les liens inter-catégoriels déjà mentionnés (*performance*/*évaluer*, *mousson*/*tropical*, *thérapeutique*/*guérir*, etc.), on peut relever un cas d'antonymie (*sécurité*/*dangers*, ex. d) ainsi que des liens sémantiques plus lâches (par exemple *société*/*identité*, ex. a). Ces observations encouragent à aborder les reformulations de requêtes par le biais de la sémantique distributionnelle. Nous présentons dans ce qui suit la base distributionnelle utilisée et son croisement avec les paires de requêtes extraites.

3.2 La base distributionnelle utilisée : les Voisins de Revues.org

La base de voisins distributionnels utilisée dans cette étude a été construite à partir de documents francophones provenant de la collection interrogée par le moteur de recherche d'OpenEdition. Plus particulièrement, il s'agit d'articles de sciences humaines issus de revues en ligne éditées par Revues.org. Le corpus constitué fait environ 253 millions de mots. Il a été soumis à une chaîne de traitement impliquant :

– une analyse syntaxique (analyseur Talismane, (Urieli et Tanguy, 2013)) ;
– une analyse distributionnelle (module Upéry, (Bourigault, 2002)).

L'analyse distributionnelle effectuée s'appuie donc sur des contextes syntaxiques : un mot peut constituer un contexte pour un autre mot s'il est rattaché à lui par une relation de dépendance syntaxique (objet, sujet, modifieur) ; ces relations de dépendance peuvent être labellisées par une préposition (« à », « de », « sur », etc.). Deux types de rapprochements sont opérés :

– des rapprochements de prédicats, c'est-à-dire de mots gouverneurs de la relation syntaxique, auxquels cette dernière est accolée ; par exemple, *évaluer_obj* et *mesurer_obj* sont voisins car ces deux prédicats partagent des contextes tels que *impact*, *dangerosité*, etc. ;
– des rapprochements d'arguments, c'est-à-dire de dépendants syntaxiques ; par exemple *social* et *culturel* sont voisins car ils partagent des contextes tels que *tâche_de*, *favoriser_obj*, etc.

La similarité des distributions (c'est-à-dire des ensembles de contextes d'apparition) de deux mots est évaluée par le score de Lin (Lin, 1998). Dans le cadre de ce travail, nous considérons que deux mots constituent une paire de voisins lorsque leur score de Lin est supérieur à un seuil fixé à 0,1.

3.3 Croisement des paires de requêtes et des voisins

Afin de comparer les relations sémantiques mises en jeu dans les reformulations de requêtes et celles calculées par l'analyse distributionnelle du corpus, nous nous sommes appuyés sur deux jeux de données.

Pour le premier, nous avons exclu les corrections orthographiques, les reformulations portant sur des mots vides et les mots incorrects (*cf.* section 3.1). Nous avons également supprimé les doublons : par exemple, bien qu'apparaissant dans plusieurs paires de requêtes (*cf.* exemples (i) et (ii) ci-dessous), les substitutions *marketing*/*communication* et *enseignement*/*éducation* ne sont prises en compte qu'une seule fois (seule la première paire de requêtes est conservée).

 © ATALA

(i) `marketing solidaire web 2 0` → `communication solidaire web 2 0`
`communication luxe` → `marketing luxe`
`marketing humanitaire` → `communication humanitaire`

(ii) `enseignement en milieux hospitaliers` → `éducation en milieux hospitaliers`
`enseignement des détenus mineurs` → `éducation des détenus mineurs`

Pour le second jeu, nous avons également supprimé les reformulations portant sur des mots absents de la base distributionnelle utilisée. Les éléments filtrés pour la constitution des deux jeux de données sont récapitulés dans le tableau 4.

Élements filtrés	Exemples		Nb
Mots-vides	`et` → `en`	(tab. 3 ex. i)	563
Corrections	`hydroterhmales` → `hydrothermales`	(tab. 3 ex. k)	5155
Mots inconnus	`iécoles` → `institutions`	(tab. 3 ex. g)	1727
Doublons	`marketing` → `communication` (x3)	(i)	399
	⇒ Premier jeu de données : 3481 paires de requêtes		

Mots absents de la base	`résistance quinine` → `résistance chloroquine`	709
	⇒ Second jeu de données : 2772 paires de requêtes	

TABLE 4 – Récapitulatif des éléments filtrés
Point de départ : 11321 paires de requêtes

4 Résultats et analyse

Le tableau 5 montre les résultats obtenus en croisant les deux jeux de données décrits dans la section 3.3 avec les voisins distributionnels.

	Paires de requêtes	Voisins	
Jeu 1	3481	1641	47%
Jeu 2	2772	1641	59%

TABLE 5 – Résultats : paires de mots substitués voisins

Ces résultats montrent tout d'abord l'intérêt de notre démarche pour les deux directions évoquées en introduction :

- le recouvrement entre les voisins distributionnels et le premier jeu de requêtes montre que la sémantique distributionnelle permet d'appréhender efficacement les relations de sens impliquées dans les reformulations de requêtes, ce qui laisse envisager d'utiliser le critère du voisinage distributionnel pour aider au typage d'autres schémas de reformulations ;
- les données choisies – qui exhibent une large gamme de relations sémantiques, sont adaptées au corpus et sont produites par des utilisateurs dans une situation concrète de recherche d'information – apparaissent comme une référence idéale pour l'évaluation (en rappel) d'une base distributionnelle. Le recouvrement observé avec le second jeu de données (59%) est supérieur à ce qui a pu être observé en confrontant des bases distributionnelles à des *gold standard* (Morlane-Hondère et Fabre, 2012).

 © ATALA

Nous avons par ailleurs croisé la présence du lien dans la base distributionnelle et différents paramètres, et avons trouvé une liaison significative avec la position des mots substitués, qui ont davantage tendance à présenter un lien de voisinage lorsque leur position est identique d'une requête à l'autre que lorsqu'elle change (test du χ^2, $p < 10^{-7}$). Cet effet est toutefois marginal (ϕ=0.1) : 61% des mots substitués à la même position sont des voisins, contre 44% si la position change.

Notons que l'expérimentation menée est adaptée à une évaluation en rappel des voisins distributionnels, mais pas en précision, du fait du manque d'exhaustivité des données : le fait qu'une paire de mots ne soit pas exploitée dans les reformulations ne signifie pas qu'elle ne présente pas de relation sémantique pertinente. Il est toutefois important de mettre en perspective le rappel obtenu afin de déterminer s'il n'est pas « gonflé » par une surgénération de voisins (il est facile d'obtenir un rappel parfait, si la précision de la base distributionnelle est quasi-nulle). En effet, le nombre de voisins des mots impliqués dans les reformulations est important (2191 en moyenne par mot). Dans cet objectif, nous avons vérifié le rang auxquel apparaît, pour chaque mot donnant lieu à une substitution, la relation exploitée dans cette substitution parmi tous ses voisins (ordonnés suivant le score de Lin décroissant). Par exemple, dans les reformulations `web documentaire` → `web reportage` ou `les jeux didactiques` → `les jeux pedagogiques`, *reportage* et *pédagogique* sont premiers parmi les voisins de, respectivement, *documentaire* et *didactique*. Par contre, dans la reformulation `espace urbain` → `aménagement urbain`, *aménagement* n'est que le 1346e voisin d'*espace* (sur 3181 voisins). L'analyse des rangs montre que pour la moitié des reformulations, le mot ajouté apparaît dans les premiers 10% de voisins du mot remplacé, ce qui montre que le fort rappel observé repose surtout sur une proximité distributionnelle importante et ne peut être imputé à une simple surgénération de voisins.

Si une proportion de 59% peut être considérée comme importante pour des tâches de confrontation de l'analyse distributionnelle à des données réelles, elle est toutefois loin d'être pleinement satisfaisante eu égard aux conditions optimales de l'expérimentation menée. On peut alors se demander quels sont les éléments impliqués dans le « silence » observé, du point de vue des mécanismes mis en jeu dans le calcul distributionnel – des mots auraient-ils pu être rapprochés mais ne l'ont pas été ? – et de la nature des données exploitées – l'hypothèse de continuité sémantique entre les deux requêtes doit-elle être mitigée ? (ce que suggère la corrélation avec la position des mots substitués), certaines relations sont-elles moins captées ? Cela nous a amenés à examiner plus en détails deux aspects : (a) les contextes distributionnels des mots substitués, qu'ils soient voisins ou non voisins (sous-section 4.1) ; (b) la nature des relations sémantiques captées ou non par le voisinage distributionnel (sous-section 4.2).

4.1 Contextes distributionnels

Nous avons mis en place plusieurs procédures d'observation visant à expliciter les mécanismes distributionnels permettant de rapprocher les termes substitués dans les requêtes.

4.1.1 Identification des contextes à l'origine des relations de voisinage distributionnel

Dans un premier temps, pour en avoir un aperçu global, nous avons regroupé tous les contextes syntaxiques communs à deux mots substitués, sur l'ensemble des 2772 paires de requêtes pour

 © ATALA

lesquelles les deux mots substitués sont présents dans la base distributionnelle. Au total, 24563 contextes syntaxiques différents ont été identifiés, que ceux-ci aient été suffisants ou non pour que les mots substitués soient déclarés comme voisins.

Nous avons calculé, pour chaque contexte ainsi identifié, le nombre de couples de mots substitués qu'il a permis de rapprocher. Le tableau 6 présente les contextes les plus fréquents. Nous y avons distingué les prédicats (à gauche) des arguments (à droite),

Prédicat	Nb. de couples
pouvoir_suj	1870
faire_suj	1634
devoir_suj	1471
concerner_obj	1421
se agir_de	1372
constituer_suj	1353
histoire_de	1258
devenir_suj	1247
forme_de	1207
sembler_suj	1206
apparaître_suj	1198
question_de	1197
mettre_obj	1187
aller_suj	1176
faire_de	1159
cas_de	1158
faire_obj	1139
rester_suj	1129
parler_de	1124
type_de	1068

Argument	Nb. de couples
tout	1441
tel	1352
nouveau	1338
autre	1272
politique	1267
seul	1144
français	1033
premier	1028
grand	1010
même	945
propre	935
page	931
différent	874
véritable	848
social	843
ancien	833
pays	819
nombreux	797
important	790
présent	745

TABLE 6 – Liste des contextes les plus fréquemment communs aux deux mots substitués

Comme on peut le voir, si certains contextes sont très génériques (sujets des verbes modaux, modifieurs *tout* ou *tel*), on y retrouve rapidement des situations spécifiques à un corpus de SHS, que ce soit par les thématiques abordées (*politique*, *histoire_de*, *français*, etc.) ou par les formes présentatives (*question_de*, *constituer_obj*, etc.).

La présence de l'argument *page* dans cette liste est un des artefacts liés à l'analyse syntaxique automatique du corpus : les syntagmes nominaux du type *page X* sont considérés comme étant des appositions, et donc rattachés à une très grande variété de noms. Par exemple, dans l'extrait « *Dans les temps modernes (page 36) [...]* », *page* est à tort considéré comme un modifieur du nom *temps* par Talismane [3].

4.1.2 Recouvrement des contextes distributionnels avec le contexte des requêtes

Dans un second temps, nous avons regardé spécifiquement si les contextes distributionnels communs se retrouvaient également dans le texte commun aux deux requêtes consécutives.

3. Qui est un analyseur statistique qui, à sa décharge, n'avait jamais rencontré ce type de configuration dans son corpus d'apprentissage...

© ATALA

Autrement dit, nous avons identifié les cas où les mots inchangés de la paire de requêtes qui forment une reformulation sont, dans la base distributionnelle, des éléments contextuels qui permettent de rapprocher les deux mots substitués.

Par exemple, dans la paire de requêtes :
`ligne aérienne internet` → `compagnie aérienne internet`
l'adjectif *aérien* est un contexte commun à *ligne* et à *compagnie* dans la base : il y a donc recouvrement des contextes entre le corpus et la requête.

Il se trouve que 27% des paires de requêtes étudiées présentent cette caractéristique avec au moins un mot de la requête correspondant à un contexte distributionnel. On y trouve à la fois des contextes très génériques comme *histoire_de* dans le couple
`histoire de l'ethnologie` → `histoire de l'anthropologie`
ou *femme_mod* dans
`femme antique` → `femme romaine`,
mais il peut s'agir également de structures très figées comme *caritatif* dans
`oeuvres caritatives` → `organisations caritatives`
ou encore *pénal* dans
`droit pénal` → `procédure pénale`.

Dans certains cas (10% des reformulations pour lesquelles le voisinage distributionnel n'a pas été identifié), nous avons pu observer la présence d'un contexte commun très spécifique, qui n'a pas permis à lui seul de permettre le rapprochement des deux termes comme voisins. C'est le cas du prédicat *mariage_mod* pour la paire de requêtes
`mariage forcé` → `mariage arrangé`
ou encore de *culture_mod* pour
`culture vivrière` → `culture maraîchère`.

Ces premières observations nous confirment le rôle central du corpus sur lequel la base de voisins est construite, mais nous amènent aussi à requestionner le mode de calcul actuel (et notamment le seuillage) de la notion de voisinage.

4.2 Relations sémantiques

Nous avons cherché à caractériser les différences sur le plan des relations sémantiques entre les deux séries de résultats : les couples de mots substitués qui sont des voisins distributionnels et ceux qui n'en sont pas. Pour cela, nous avons annoté deux échantillons de 200 couples, en leur attribuant les étiquettes qui sont mentionnées dans le tableau 7.

Cette annotation ne peut fournir que des indications très sommaires. On sait en effet la difficulté à annoter des relations sémantiques hors contexte (Adam, 2012). Nous pouvons néanmoins dégager deux types d'observation.

Tout d'abord, ces chiffres donnent un aperçu de la nature des liens qui sont mobilisés dans les reformulations. Ils confirment la prépondérance des liens sémantiques non classiques. Les cinq premières relations listées dans le tableau concernent seulement 1/3 des mots substitués, l'hyperonymie (généralisation ou spécification) et la cohyponymie (mouvement parallèle) étant les relations dominantes. On voit en effet que les relations dominantes entre mots substitués sont plus difficiles à caractériser et témoignent d'un saut conceptuel plus important pour atteindre

 © ATALA

Nature du lien sémantique	Exemples	V (%)	non V (%)	tous (%)
synonymie	emploi-travail, mensonger-trompeur	11	3	7
lien morphologique	urbanisme-urbanisation, sacralité-sacré	3,5	6	4,75
antonymie	exclusion-intégration, sécurité-danger	2	0,5	1,25
hyperonymie	mariage-alliance, drogue-cannabis	13,5	6,5	10
méronymie	Angola-Afrique, puce-mémoire	0,5	3	1,75
cohyponymie	guadeloupe-martinique, reportage-documentaire	9	7,5	8,25
champ lexical	enfant-école, politique-engagement	25	31	28
autre lien	urbanisation-attractivité, attention-motivation	28,5	22	21,75
absence de lien	médiation-mine, banque-antécédent	7	20,5	13,75

TABLE 7 – Relations sémantiques identifiées dans les résultats substitués

un mot du même champ lexical voire un lien associatif plus lâche. On peut s'étonner de trouver une proportion non négligeable de mots substitués sans lien sémantique évident. Cette situation correspond à plusieurs cas de figure illustrés ci-dessous. Les requêtes (i) et (ii) correspondent à un double mouvement d'effacement et d'ajout, plutôt qu'à une substitution (on constate d'ailleurs un changement de position entre le mot originel et le mot substitué). Les exemples (iii), (iv) et (v) montrent des cas de redénomination qui rendent difficiles la mise en correspondance des composants substitués (*américanisation* et *exception* présentent une relation de contraste dans ce seul contexte très particulier). Enfin, les requêtes (v) et (vi) illustrent des cas de modification importante dans la désignation de la cible de la requête.

(i) `Maurice Genty` → `Genty contoverses`

(ii) `extériorisation logistique` → `risques extériorisation`

(iii) `Web 2.0` → `ville 2.0`

(iv) `jeux olympiques 2008` → `jeux olympiques Pékin`

(v) `exception culturelle` → `américanisation culturelle`

(vi) `mimétisme alimentation` → `media alimentation`

(vii) `art engagé` → `art osé`

Ces premières observations confirment la nécessité d'identifier des liens sémantiques variés lorsqu'il s'agit de mettre en place des procédures visant à accompagner les procédures de reformulation. Ces chiffres montrent aussi que l'ensemble des relations sont captées, dans des proportions néanmoins très variables, par le critère de proximité distributionnelle.

On peut par ailleurs relever les différences les plus nettes entre les deux échantillons. Elles concernent la part de synonymes (22 des 28 synonymes ont été repérés par le voisinage distributionnel), et la proportion de couples sans relation sémantique : ces deux chiffres attestent du rôle de filtre du voisinage distributionnel, qui repère les liens sémantiques les plus étroits et écarte les mots sémantiquement éloignés. Ce filtre n'est pourtant pas totalement efficace. Le non repérage de certains synonymes s'explique en particulier par des fréquences très déséquilibrées entre les deux mots (ex : *litige*/*conflit*, *herméneutique*/*interprétation*). De même, le voisinage distributionnel contribue fortement à filtrer les couples sans lien sémantique identifiable mais il ne les exclut pas totalement. Ainsi, les mots *médiation* et *mine* sont rapprochés parce qu'ils peuvent tous deux figurer en complément de certains noms très généraux (*travail*, *champ*, *produit*, *service*, *développement*, *système*...). La plupart de ces couples sont eux-mêmes composés de mots très généraux, aux contextes par conséquent peu discriminants (*réseau*/*genre*, *science*/*gouvernance*, *démocratie*/*pensée*, *type*/*technique*, *monde*/*mouvement*...).

 © ATALA

5 Conclusion et perspectives

Cette étude a montré que le critère distributionnel permet de détecter des types variés de proximité sémantique, qui sont mobilisés dans l'activité de reformulation que pratiquent des utilisateurs d'un moteur de recherche : les opérations de substitution sémantique qui sont utilisées dans les requêtes font appel à des relations de sens multiples, et majoritairement à des relations lâches qui sont bien détectées par le critère du voisinage distributionnel (cohyponymes, mots appartenant au même champ lexical, liens très contextualisés). L'analyse des résultats a permis par ailleurs de mettre au jour les faiblesses de l'analyse distributionnelle sur deux points principaux : la surgénération de voisins à partir de contextes partagés trop généraux ; la non détection de mots proches qui ne sont liés que par une série très limitée de contextes, voire un contexte unique (comme *forcé* et *arrangé* modifiant le seul mot *mariage*).

Si l'on se place dans une perspective de recherche d'information (avec un objectif à long terme d'assistance de l'utilisateur ou de détection de reformulation), on peut considérer que des ressources du type des bases distributionnelles sont effectivement adaptées. Toutefois, leur mode de calcul et de projection doit être spécifiquement affiné pour lutter contre le bruit que génère ce type de méthode. On peut également envisager d'exploiter les informations contextuelles des requêtes (ne serait-ce que l'ensemble des mots de celles-ci, ou l'historique des requêtes d'un utilisateur) pour combler les lacunes des calculs distributionnels.

Remerciements

Ce travail s'inscrit dans le cadre du projet ANR CAAS (*Contextual Analysis and Adaptative Search*), programme CONTINT (2010-2014), coordonné par Josiane Mothe (IRIT).

Merci à Marin Dacos, à Patrice Bellot et à toute l'équipe du CLEO pour nous avoir permis d'accéder à leurs logs de requêtes ainsi qu'à leur corpus.

Merci à Simon Leva (CLLE) et à Nicolas Faessel (IRIT) d'avoir filtré, croisé et dépouillé les différents logs d'accès.

Merci à Franck Sajous (CLLE-ERSS) pour son travail de longue haleine sur la construction des bases distributionnelles.

Références

ADAM, C. (2012). *Voisinage lexical pour l'analyse du discours*. Thèse de doctorat, Université de Toulouse.

ANGUIANO, E., DENIS, P. *et al.* (2011). Fredist : Automatic construction of distributional thesauri for french. *In Actes de TALN*, pages 119–124.

BARONI, M. et LENCI, A. (2010). Distributional memory : A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

BOURIGAULT, D. (2002). UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *In Actes de TALN*, pages 75–84, Nancy.

 © ATALA

GREFENSTETTE, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Dordrecht : Kluwer Academic Publishers.

HUANG, J. et EFTHIMIADIS, E. N. (2009). Analyzing and evaluating query reformulation strategies in web search logs. *In Proceedings of the 18th ACM conference on Information and knowledge management*, pages 77–86. ACM.

JANSEN, B., SPINK, A. et SARACEVIC, T. (2000). Real life, real users, and real needs : a study and analysis of user queries on the web. *Information Processing and Management*, 36:207–227.

JANSEN, B. J., BOOTH, D. L. et SPINK, A. H. (2009). Patterns of query reformulation during web searching. *American Society for Information Science and Technology Journal*, 60(7):1358–1371.

LAFOURCADE, M. et JOUBERT, A. (2010). Construction de l'arbre des usages nommés d'un terme dans un réseau lexical évolutif. *Actes des Journées internationales d'Analyses statistiques des Données Textuelles (JADT'10)*.

LEVA, S. (2013). Les sessions de recherche comme contexte des requêtes. *In Actes de l'atelier sur la Contextualisation des Messages Courts, dans le cadre de EGC*, pages 1–13, Toulouse.

LIN, D. (1998). An information-theoretic definition of similarity. *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, Madison.

MORLANE-HONDÈRE, F. et FABRE, C. (2012). Le test de substituabilité à l'épreuve des corpus : utiliser l'analyse distributionnelle automatique pour l'étude des relations lexicales. *In Actes du Congrès Mondial de Linguistique Française*.

MORRIS, J. et HIRST, G. (2004). Non-classical lexical semantic relations. *In Proceedings of the HLT Workshop on Computational Lexical Semantics*, pages 46–51, Boston.

PADÓ, S. et LAPATA, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

PICTON, A., FABRE, C. et BOURIGAULT, D. (2008). Méthodes linguistiques pour l'expansion de requêtes. Une expérience basée sur l'utilisation du voisinage distributionnel. *Revue Française de Linguistique Appliquée*, XIII(1):83–96.

SAJOUS, F., HATHOUT, N. et CALDERONE, B. (2013). Glàff, un gros lexique à tout faire du français. *In Actes de TALN*.

URIELI, A. et TANGUY, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur talismane. *In Actes de TALN*.

van der PLAS, L. et TIEDEMANN, J. (2008). Using lexico-semantic information for query expansion in passage retrieval for question answering. *In Coling 2008 : Proceedings of the 2nd workshop on Information Retrieval for Question Answering*, pages 50–57. Association for Computational Linguistics.

WEEDS, J. (2003). *Measures and Applications of Lexical Distributional Similarity*. Thèse de doctorat, University of Sussex.

© ATALA

Interactive visualizations of Semantic Vector Spaces for lexicological analysis

Thomas Wielfaert Kris Heylen Dirk Speelman
QLVL, University of Leuven, Faculty of Arts, Blijde-Inkomststraat 21/3308, B-3000 Leuven, Belgium
{thomas.wielfaert, kris.heylen, dirk.speelman}@arts.kuleuven.be

RÉSUMÉ

Visualisations interactives des espaces vectoriels sémantiques pour l'analyse lexicologique

Dans le domaine de la Linguistique Computationnelle, les modèles sémantiques distributionnels sont devenus les piliers pour modéliser la sémantique lexicale à grande échelle. La modélisation distributionnelle présente également un grand potentiel pour la recherche dans le domaine de la Linguistique proprement dite. Elle permet aux linguistes de baser leurs analyses sur des grandes quantités de données authentiques et d'élargir ainsi considérablement leur base empirique, ainsi que de détecter des motifs sémantiques potentiellement intéressants. Cependant, jusqu'à présent, il y a eu relativement peu d'applications, principalement en raison de la complexité technique et de l'absence d'une interface conviviale permettant aux linguistes d'explorer les résultats obtenus. Dans cet article, nous proposons une visualisation interactive d'une matrice de similarité distributionnelle basée sur le Positionnement Multidimensionnel. Nous présentons un prototype d'un outil de visualisation que nous avons construit en Processing [1]. Il permettra d'ouvrir de nouvelles voies pour l'analyse visuelle des modèles au niveau de l'occurrence (*token*) et nous l'appliquons à une petite étude de cas d'un mot polysémique néerlandais.

ABSTRACT

Within Computational Linguistics, distributional models of semantics have become the mainstay of large-scale modelling of lexical semantics. Distributional modelling also holds a large potential for research in Linguistics proper : It allows linguists to base their analysis on large amounts of usage data, thus vastly extending their empirical basis, and makes it possible to detect potentially interesting semantic patterns. However, so far, there have been relatively few applications, mainly because of the technical complexity and the lack of a linguist-friendly interface to explore the output. In this paper, we propose an interactive visualization of a distributional similarity matrix based on Multi-Dimensional Scaling. We present our prototype for a visualization tool built in Processing [1] which opens up new possibilities for the visual analysis of token-based models and apply it to a small case study of a Dutch polysemous word.

MOTS-CLÉS : visualisation, sémantique distributionnelle, espaces vectoriels sémantiques, niveau d'occurence, positionnement multidimensionnel, sémantique lexicale, lexicographie, néerlandais.

KEYWORDS: visualization, distributional semantics, semantic vector spaces, token-level, multidimensional scaling, lexical semantics, lexicography, Dutch.

1. http ://www.processing.org

 © ATALA

1 Introduction

Distributional approaches to lexical semantics model a word's meaning in terms of the contexts it appears in. In Computational Linguistics this idea has been implemented in the form of high-dimensional co-occurrence matrices that can be manipulated with linear algebra. Under the heading of Vector Space Models (VSMs), Semantic Vector Spaces (SVSs) or Word Spaces, distributional models have since the middle of the 1990s, become the mainstay in statistical NLP for the large-scale analysis of lexical semantics (see Turney and Pantel (2010) for an overview). They have been applied to different tasks like Thesaurus Extraction, Word Sense Disambiguation or Lexical Entailment, and are traditionally evaluated in terms of precision, recall and F-score against a gold standard. On the other hand, distributional approaches have been used in Linguistics proper and Applied Linguistics to study lexical semantics for over half a century in the form of collocational analysis. However, these studies analyse co-occurrence frequencies on a much smaller scale than the computational Vector Space Models and tend to look at individual words in their most typical usage contexts. The identification and interpretation of more general patterns is typically left to the linguists themselves. The possibilities of Vector Space Models for large-scale semantic modelling and automatic pattern finding in Big Data have not been widely adopted in theoretical and applied linguistic subdisciplines. However, Vector Space implementations could greatly contribute to the analysis of lexical semantics in lexicology (the linguistic study of word meaning), provided that they are made accessible to lexicologists through an interactive visualization of their output that also allows easy access to the original input data and that enables the linguists to gauge effect of different parameter settings in the models. In this paper, we report on work in progress in the development of such a visualization and the features and functionalities that we added in response to the needs experienced in a practical case study. As such, the work in this paper is situated on the cross roads of Distributional Semantic Modelling in Computational Linguistics, Visual Analytics and (Applied) Linguistics. Although the creation of a tool for linguistic analysis is our first priority, such a tool could also allow computational linguists to tap into the expert knowledge of lexicologists and lexicographers to perform an in-depth, qualitative evaluation of their models that can complement the traditional precision and recall scores obtained on standard tasks.

Our visualization of Semantic Vector Spaces was guided by two traditional subtasks in lexicology : (1) identifying the different senses of a word in a large set of attestations of that word, and (2) charting the variation in word meaning and word choice between different varieties of the same language. Consequently, our vector spaces are situated in the Word Sense Induction paradigm and have vector representations for individual occurrences (tokens) of a word. In order to study lexical *variation*, word tokens are represented by vectors from multiple corpora that are representative of different varieties of the same language. Differences in word choice for the same concept are analysed by simultaneously modelling tokens from a set of near-synonymous words rather than from one word type at the time. This way we recreate the typical input for a lexicological analysis, viz. a long list of concordances of a set of near-synonyms, in which the lexicologist has to identify structure in terms of word senses and regional patterns. The idea is now that a token-level semantic vector space can help the lexicologist to discern the structure and patterns quicker and more easily through statistical modelling of word distributions. Importantly, the aim is NOT to provide a ready-made analysis of the semantic structure, but rather a starting point for analysis by a human expert. That expert can then interact with and improve upon the discovered semantic structure by inspecting the tokens' concordances. It is therefore important

 © ATALA

that the lexicologist does not experience the semantic vector space as a black box, but rather can see the effect of different parameter settings and play with them. Such more direct assessments of the semantics captured by distributional models is also desirable from a Computational Linguistic point of view : Baroni and Lenci (2011) point out that "To gain a real insight into the abilities of DSMs (*Distributional Semantic Models*, A/N) to address lexical semantics, existing benchmarks must be complemented with a more intrinsically oriented approach, to perform direct tests on the specific aspects of lexical knowledge captured by the models".

For this interactive and intrinsic assessment we first construct a number of token vector spaces where we vary parameters like weighting and context size. Next, we make two dimensional visualizations of the semantic structure that is captured by the high-dimensional vector spaces and we present these visualization in an interactive chart where the lexicologist can inspect the concordances and track the effect of the different parameters. The two dimensional visualization relies on a Multidimensional scaling analysis of the token-by-token similarity matrix that is calculated from the underlying token vectors. For the interactive functionality of the charts, we have tried out different packages including the Python Image Library and Google Motion Charts. As a response to their short-comings in desired functionality, we developed our own tool in the Java-based programming language Processing. The usability and functionality of the tool was tested in a case study where we look at a set of near-synonyms in Belgian and Netherlandic Dutch that refer to the concept COMPUTER SCREEN.

The rest of this paper is structured as follows. In the next section we discuss related approaches in distributional semantics and visual analytics. Section 3 introduces our corpus and our implementations of a token-level semantic vector space. In Section 4 we discuss the different visualizations we experimented with. In Section 5 we present our interactive visualization tool and discuss how its functionality was adapted in response to needs in the case study. Finally, Section 6 wraps up with conclusions and prospects for future research.

2 Related work

A number of approaches have applied distributional semantic models to the linguistic study of lexical variation. Diachronic variation, or meaning change, has been modelled on a type and a token-level. Gulordava and Baroni (2011), Cavallin (2012), Cook and Hirst (2011) try to detect new meanings of word types in general. Cook and Stevenson (2010) look specifically at the change in semantic orientation of word types. On a token-level, Sagi et al. (2009) aims to model semantic generalisation and specialisation processes. Rohrdantz et al. (2011) and Lau et al. (2012) try to identify tokens with new meaning using topic models. Regional variation in word semantics is modelled by Peirsman et al. (2010).

The visualization of distributional semantic models has most notably been undertaken by Rohrdantz et al. (2011) They explicitly aim to combine distributional semantics and visual analytics to track changes in word meaning. More specifically, they train a topic model on text snippets in which words occur that have undergone recent semantic change like *mouse* or *surf*. They then plot for each topic how the number of occurrences changed over time. This visualizes how topics that represent newer meanings, have more high-scoring tokens in more recent years.

Another approach to visualize high-dimensional semantic space was presented by Kievit-Kylar

 © ATALA

and Jones (2012). In their "Word-2-Word" visualization tool, words are represented as nodes and word similarities are represented by directed edges between these nodes. The software implements 22 common similarity-based metrics, both corpus- and WordNet-based, and the visualizations allow direct user input. They present various case studies varying from an artificial language, over the CHILDES corpus to a visual analysis of Bush and Obama State of the Union Adresses.

The most direct precursor of our approach is Heylen et al. (2012) who propose a token-level, interactive visualization of the regional variation in the semantics of Dutch near-synonyms. They present an interactive scatter plot using Google Chart Tools, but point out that these charts are not designed to display larger chunks of texts that are needed for informative concordances. It is this approach whose interactive visualization we try to optimize in this paper.

3 Corpus and Token-level SVS

The corpus for our study consists of Dutch newspaper materials from 1999 to 2005. For Netherlandic Dutch, we used the 500M words Twente Nieuws Corpus (Ordelman, 2002) [2], and for Belgian Dutch, the Leuven Nieuws Corpus (aka Mediargus corpus, 1.3 million words [3]). The corpora were automatically lemmatized, part-of-speech tagged and syntactically parsed with the Alpino parser (van Noord, 2006).

Token-level SVSs have been developed for modelling word meaning in context for different tasks (see Dinu et al. (2012) for an overview). The token-level SVS we use here is a fairly basic implementation and constitutes an adaptation of the early approach proposed by Schütze (1998). He models the semantics of a token as the frequency distribution over its so-called second order co-occurrences. These second-order co-occurrences are the type-level context features of the (first-order) context words co-occurring with the token. This way, a token's meaning is still modelled by the "context" it occurs in, but this context is now modelled itself by combining the type vectors of the words in the context. This higher-order modelling is necessary to avoid data-sparseness : any token only occurs with a handful of other words and a first-order co-occurrence vector would thus be too sparse to do any meaningful vector comparison. Note that this approach first needs to construct a type-level SVS for the first-order context words that can then be used to create a second-order token-vector.

In our study, we therefore first constructed a type-level SVS for the 573,127 words in our corpus with a frequency higher than 2. Since the focus of this study is visualization rather than finding optimal SVS parameter settings, we initially chose a number of settings that gave good results in our previous studies Peirsman et al. (2008); Heylen et al. (2008); Peirsman et al. (2010). For the context features of this SVS, we used two bag-of-words approaches, one with a window of 4 to the left and right around the targets and one with a window of 7 words on both sides of the context word. Furthermore, we experimented with two different weighting schemes : Positive Pointwise Mutual Information (PPMI), where negative PMI values are set to zero on the one hand and an adapted version of Log Likelihood Ratio (LLR) on the other. The context feature set was restricted to the 5430 words, that were among the 7000 most frequent words in the

2. Publication years 1999 up to 2002 of *Algemeen Dagblad*, *NRC*, *Parool*, *Trouw* and *Volkskrant*

3. Publication years 1999 up to 2005 of *De Morgen*, *De Tijd*, *De Standaard*, *Het Laatste Nieuws*, *Het Nieuwsblad* and *Het Belang van Limburg*

© ATALA

corpus, (minus a stoplist of 34 high-frequent function words) AND that occurred at least 50 times in both the Netherlandic and Belgian part of the corpus. The latter was done to make sure that Netherlandic and Belgian type vectors were not dissimilar just because of topical bias from proper names, place names or words relating to local events.

In a second step, we took a random sample of 100 Netherlandic and a 100 Belgian newspaper issues from the corpus and extracted all occurrences of 476 nouns in 218 semi-automatically constructed Dutch synsets by Ruette et al. (2012). For each occurrence, we built a token-vector by averaging over the type-vectors of the words in a window of 10 words to the left and right of the token. We experimented with two averaging functions. In a first version, we followed Schütze (1998) and just summed the type vectors of a token's context words, normalizing by the number of context words for that token :

$$\vec{o_i^w} = \frac{\sum_{j \in C_i^w}^{n} \vec{c_j}}{n}$$

where $\vec{o_i^w}$ is the token vector for the ith occurrence of noun w and C_i^w is the set of n type vectors $\vec{c_j}$ for the n context words in the window around that i^{th} occurrence of noun w. However, this summation means that each first-order context word has an equal weight in determining the token vector. Yet, not all first-order context words are equally informative for the meaning of a token. In a sentence like "While walking to work, the teacher saw a dog barking and chasing a cat", $bark$ and cat are much more indicative of the meaning of dog than say $teacher$ or $work$. In a second, weighted version, we therefore increased the contribution of these informative context words by using the first-order context words' PMI or LLR values with the noun in the synset. PMI and LLR can be regarded as measures for informativeness and target-noun/context-word weights were available already from our large type-level SVS. The PMI or LLR of a noun w and a context word c_j can now be seen as a weight $pmi_{c_j}^w$ or $llr_{c_j}^w$. In constructing the token vector $\vec{o_i^w}$ for the ith occurrence of noun w with PMI as weight, we now multiply the type vector $\vec{c_j}$ of each context word with the PMI weight $pmi_{c_j}^w$, and then normalize by the sum of the pmi-weights :

$$\vec{o_i^w} = \frac{\sum_{j \in C_i^w}^{n} pmi_{c_j}^w * \vec{c_j}}{\sum_j^n pmi_{c_j}^w}$$

The same can be done with LLR as weight. The token vectors of all nouns from the same synset were then combined in a token by second-order-context-feature matrix. Note that this matrix has the same dimensionality as the underlying type-level SVS (5430). By calculating the cosine between all pairs of token-vectors in the matrix, we get the final token-by-token similarity matrix for each of the 218 synsets [4].

3.1 A polysemous case

Heylen et al. (2012) discusses the interesting case of the concept COMPUTER SCREEN in Dutch which is represented by three near-synonyms : *computerscherm*, *beeldscherm* and *monitor*. The

4. string operations on corpus text files were done with Python 2.7. All matrix calculations were done in Matlab R2009a for Linux

© ATALA

latter appeared to have interesting properties : the first meaning of *monitor* is '(computer) screen' or 'display' in both Netherlandic and Belgian Dutch. In Belgian Dutch, however, *monitor* is polysemous and can also refer to a *supervisor of youth leisure activities*, henceforth *youth leader* [5].

To demonstrate the sensitivity of Semantic Vector Spaces models we limit our case study to the *monitor* tokens. In a token-level SVS for these tokens, one can expect two distinct token clouds : one with all the *screen* tokens and a second one with the *youth leader* tokens. The latter should only contain tokens from Belgian Dutch newspapers and we expect more instances from 'popular' newspapers than from 'quality' newspapers because the *youth leader* meaning is explicitly marked as Belgian Dutch in dictionaries and is thus not regarded as Standard Dutch.

To verify whether the spatial position of a token corresponds to the actual meaning, we needed manually disambiguation of the tokens. [6] Using their context, the tokens were manually assigned to three categories : 'display', 'youth leader' and 'other'. It is explicitly not the aim to apply time consuming manual disambiguation on a larger scale. We solely use this manual approach to analyse the effects of the different parameters on the result.

4 Towards an interactive visualization

The token-by-token similarity matrix is a high-dimensional matrix, which is not very informative by itself. To visualize it in 2 or 3 dimensional representation, we need to apply a dimension reduction technique. One way to visualize a similarity matrix is to apply Multidimensional Scaling (MDS) (Cox and Cox, 2000). MDS reduces the high-dimensional space to a 2 or 3 dimensional representation while respecting the original distances in high-dimensional space in the best possible way. We apply Kruskal's non-metric Multidimensional Scaling in R using the ISOMDS function from the MASS package. The degree of faithfulness to the original distance for the MDS solution is called stress. The stress values in this case study are roughly between 20 and 25% which is quite high, but still acceptable.

Our first attempt to visualize the token clouds was an R scatter plot (see Figure 1) [7]. This type of visualization allows to quickly eyeball the different types in the token cloud, but the level of interaction is rather low which make it fairly difficult to analyse a specific token [8]. In the next step, we integrated more of meta-data, such as country and newspaper and assign a colour code for the former categories so that in addition to word type, each country and newspaper gets its own colour. Heylen et al. (2012) uses Motion Charts from the Google Chart Tool to colour-code the data points in an interactive chart using the R-package GOOGLEVIS as an interface between R and the Google Motion Charts. These interactive charts for 218 Dutch near-synonyms (Ruette et al., 2012) can be explored online as well [9]. The chart for COMPUTER SCREEN [10] in Figure 2 shows

5. A recent example from the Belgian 'popular' newspaper *Het Nieuwsblad* found in a report on a court case following an accident during a youth leisure activity : "Volgens het [H]of van [B]eroep was er onvoldoende toezicht door de monitoren." (According to the Court of Appeal there was insufficient supervision by the youth leaders.) `http://www.nieuwsblad.be/article/detail.aspx?articleid=DMF20130403_00528125`

6. Jocelyne Daems deserves our eternal gratitude for performing this drudgery task as part of her M.A. project.

7. This is the token cloud for the concept of COMPUTER SCREEN with 3 near-synonymous types : *beeldscherm*, *computerscherm* and *monitor*

8. A more interactive version of the R plot can be consulted online : `https://perswww.kuleuven.be/~u0038536/committees/data/beeldscherm/`

9. `https://perswww.kuleuven.be/~u0038536/googleVis`

10. `https://perswww.kuleuven.be/~u0038536/googleVis/wSOCC/beeldscherm.html`

© ATALA

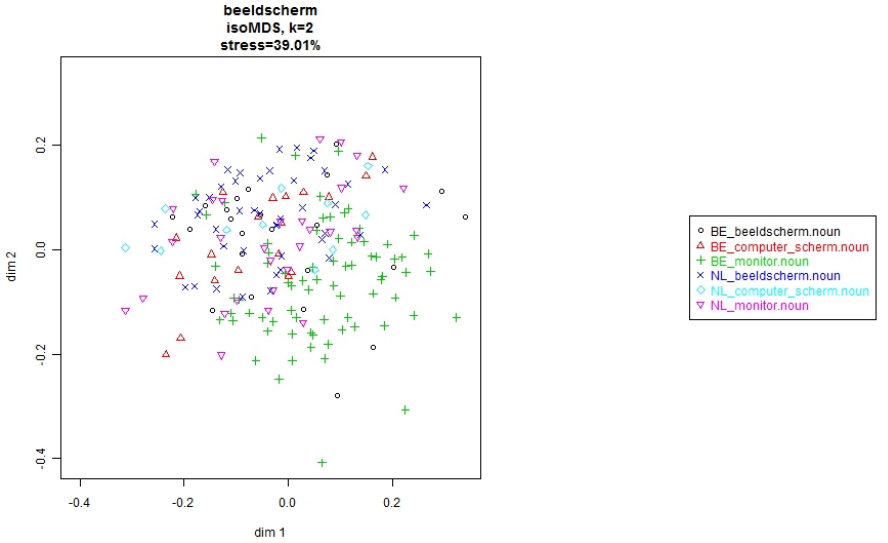


FIGURE 1 – Visualization of COMPUTER SCREEN

that the Belgian *monitor* tokens form a separate cloud on the left-hand side [11] of the plot.

5 Case study : *monitor* tokens

As we base our tests on the manually disambiguated *monitor* tokens from both corpora, we use this example as a 'touchstone' to try different parameter settings. The visualizations become more informative by displaying the underlying properties of a token in the data points of the scatter plot. Despite the adaptability of the Google Charts it is not possible to display additional (textual) information next to the plot that could enhance the interpretation of the clouds. Therefore, we propose an early version of our own visualization tool, built in Processing [12].

For the evaluation of our manually disambiguated *monitor* tokens, we colour-code the tokens based on the manually assigned senses rather by country, newspaper or word type. The red dots are the 'youth leader' tokens, the green ones represent the 'display' tokens and we use blue for other categories. For now, we can ignore the blue, other tokens because of their small number

11. The isoMDS algorithm can freely rotate the solution, so one should be careful with describing the plot in terms of left and right.

12. Processing is a Java-based programming language for quick visual applications It allows to build GUIs with a fraction of the code Java Swing would require. Processing has a Javascript mode which relies and HTML5 instead of Java. See `http://www.processing.org` and `http://www.processingjs.org` for more information.

 © ATALA

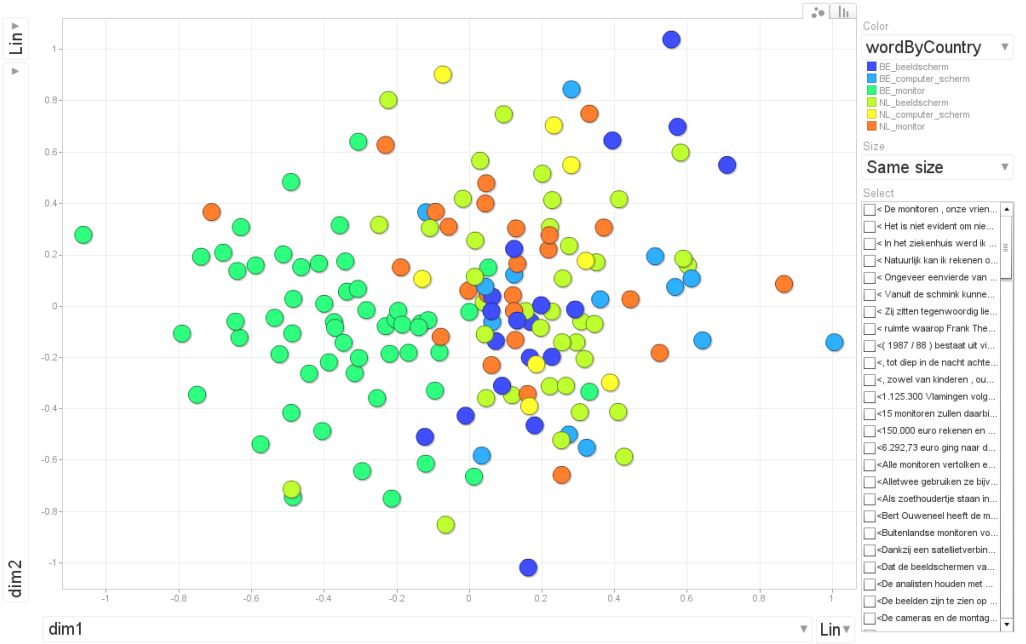


FIGURE 2 – Google Motion Chart for COMPUTER SCREEN

© ATALA

and diverse nature. Interested readers can explore the online version [13] of the plot themselves.

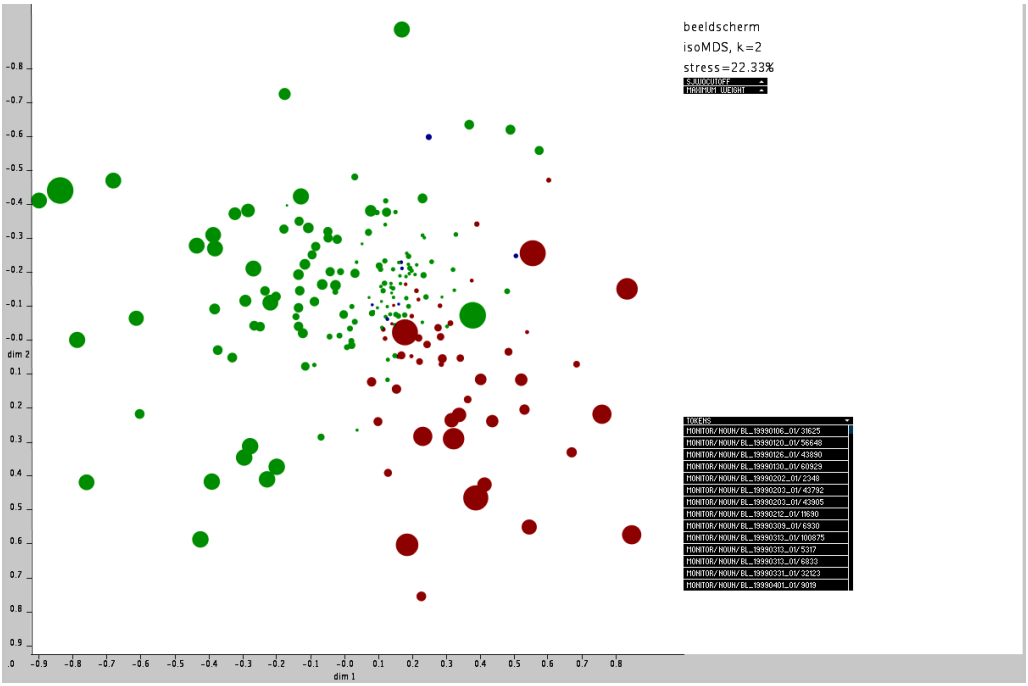


FIGURE 3 – Processing implementation of the *monitor* tokens

In Figure 3 we can see three drop down menus, right of the plot. With the first one, we can choose the dataset we want to visualize. The selection shown in Figure 3 is 'SJwoCutoff', where 'SJ' refers to the name of the sample and 'woCutoff' (without cut-off) means that we used the original weights rather than replacing values below the cut-off with a (low) standard value. For the other parameters we used our standard settings : 10-10 first-order co-occurrences window, 4-4 second-order co-occurrences window and Positive Pointwise Mutual Information for the weighting. The second drop down menu can be used to change the size of the dot depending on the weights assigned to the words within the 10-10 context window for each token. We have implemented three options : the maximum weight, the sum of weights and the absolute number of non-zero weights. Exploring these parameters allows the user to get a better insight in the impact of the weighting on the token's position.

5.1 Misclassification analysis

Analysing a token's position is particularly useful in the case of 'misplaced' tokens. In Figure 3 we can spot a very large green dot between the red ones in right half of the token space. This

13. https://perswww.kuleuven.be/~u0083608/tcViz/

© ATALA

means that a 'display' token is closer to the 'youth leader' than to the 'display' sense despite a high-weighted context word. For improving the algorithm, it is indispensable to get an insight in the factors that pushed this token out of the 'display' token cloud where it belongs. We can analyse the token properties by clicking on the dot (see Figure 4) and examine the broader context of this token. The curly brackets in the text delimit the scope of the context window while the square brackets define the context words accompanied by their respective weight.

For the selected token in Figure 4, the context word *risicobaby* [14] (risk baby) has been assigned a weight of 9.36. As the size of the dot corresponds to the maximum weight over all the context words, *risicobaby* is the major factor that puts the token in the wrong cloud. Other context words with a weight above 1 should also be considered, namely *zoeken* (search), *leggen* (lay) and *ouder* (parent). Focussing on the context words weighted above 1 is somewhat arbitrary, but the function words, which are not considered very informative in bag-of-word models, tend to have a weight below 1 with PMI weighting.

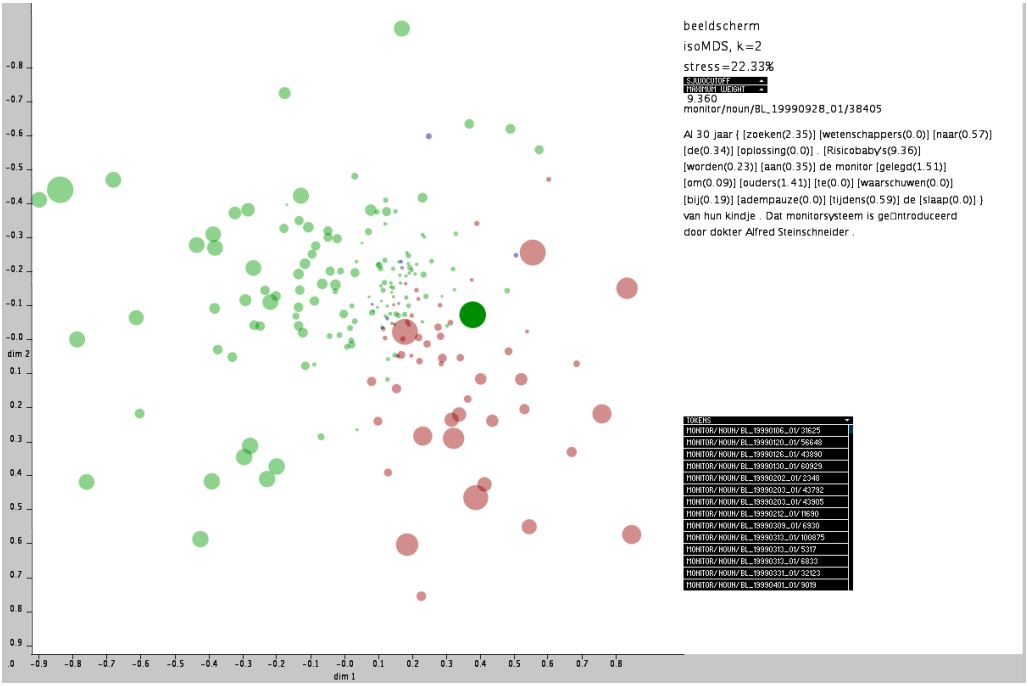


FIGURE 4 – Token cloud visualization with misclassified *risicobaby* token selected

The visualization with variable dot size allowed us to quickly eyeball misclassified tokens with misleading context words such as *risicobaby*. The word *risicobaby* has a very high weight (9.36) compared to the second most informative context word, which is *zoeken* (2.35). With a total number of 12 occurrences in our Dutch news corpora, *risicobaby* is a low-frequent word. The high weight of *risicobaby* illustrates two more general aspects. First, we use token-level SVSs because

14. *risicobaby's* is plural in the original text, but the corpora have been stemmed.

© ATALA

they are, in contrast with type-level SVSs, sensitive to polysemy of the co-occurrences and are capable of modelling subtle meaning differences. Nevertheless, our model is still sensitive to the properties of an individual type vector, i.e. *risicobaby*. Apparently, the context words of *risicobaby* overlap more with the typical context words a 'youth leader' than with those of a 'display'. Second, the large weight of *risicobaby* is an artefact of the PMI weighting ; Pointwise Mutual Information is sensitive to low-frequency events (Manning and Schütze, 1999). Low-frequent word types are thus more probable to get a high weight than high-frequent ones. In general, this property is not problematic because low-frequent words can be very informative co-occurrences. In this specific case though it associates the token with the wrong meaning and there are no other informative context words that could compensate.

We could remediate the frequency effect of PMI in two ways : use a larger context window for the weighting, for instance 7-7 instead of 4-4 and hope that a bigger window increases the number of informative context words and thus lowers the relative weight of *risicobaby*. We explored this option with the 'SJ77' dataset and we found that the weight of *risicobaby* indeed lowered (8.8), but not enough to 'push' it out of the 'youth leader' cloud. Another possible approach is to use a different association measure such as Log Likelihood Ratio instead of PMI. Again we can vary the window size or try different cut-off values. We invite the reader to compare these different parameter settings on our website.

6 Discussion

With our new tool to visualize similarity matrices we continued the work Heylen et al. (2012). Developing our own visualization tool allows us to imitate the visual appeal of the Google Chart Tools while remediating its shortcomings for text analytics. Our main goal remains to open up the black box of Semantic Vector Spaces, not only to computational linguists, but also to lexicographers and lexicologists with an interest in empirical quantitative data analysis. Of course, this is still a work in progress and there are some possible improvements.

First and foremost, it is clear that there is almost an infinite number of parameters that can be used in order to get more coherent token clouds. However, we need a more formal and efficient measure to evaluate the different solutions. We already tried a fairly simple and naive algorithm called CLUSTER QUALITY we borrowed from Speelman and Geeraerts (2009). The basic idea is that for every token, we sum the distances to all the other tokens that belong to the same manually disambiguated cluster and divide it by the distances to the tokens that belong to a different cluster. When we then average over all clusters, the result is the cluster quality value. Unfortunately, we found that this algorithm is so sensitive to outliers that the cluster quality of the real similarity matrix can not reliably be distinguished from a distance matrix created with random numbers. A possible solution is taking only the k-nearest neighbours into account. The value of k is something that needs to be determined in future research. This approach is of course only useful if we have manually disambiguated data. Obtaining these is very labour intensive and contrary to our goal to use unsupervised learning for lexical semantics. For the analysis of the larger dataset, we might have to experiment with traditional cluster analysis algorithms. The number of clusters should then ideally be decided by the user and entered as a parameter into the visualization tool.

Secondly, we need to further explore the possibilities to make the SVS's black box mechanisms

 © ATALA

more accessible. We can now visualise the weights of the first-order co-occurrences and explore some simple calculations based on these weights, but there are more underlying parameters that could be unveiled. We could for instance make the properties of the second-order co-occurrence vectors explicit.

Thirdly, creating 3D plots is still on our wish list. Not only is it visually more appealing, it could also reduce the stress of the isoMDS solution. For our *monitor* tokens, the stress values are moderate (roughly ranging between 0.2 and 0.25), but we restricted the plots to one word type in order to explore the parameter settings.

Finally, the plots do not allow real user input, mainly because the Semantic Vector Spaces are calculated in advance. There are technical and practical limits for using real time calculations that are beyond the scope of this paper. However, we could for instance let the user move wrongly positioned tokens and recalculate the MDS co-ordinates based on these corrections. The isoMDS algorithm is fast and requires little resources which makes a real time calculations feasible. This kind of feature could potentially benefit other fields which make use of high-dimensional similarity matrices.

References

Baroni, M. and Lenci, A. (2011). How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '11, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cavallin, K. (2012). Automatic extraction of potential examples of semantic change using lexical sets. In *Proceedings of KONVENS 2012 (LThist 2012 workshop)*, pages 370–377.

Cook, P. and Hirst, G. (2011). Automatic identification of words with novel but infrequent senses. In *Proceedings of the 25th Pacific Asia Conference on Language Information and Computation (PACLIC 25)*, pages 265–274, Singapore.

Cook, P. and Stevenson, S. (2010). Automatically identifying changes in the semantic orientation of words. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 28–34, Valletta, Malta.

Cox, T. F. and Cox, M. (2000). *Multidimensional Scaling, Second Edition*. Chapman and Hall/CRC, 2 edition.

Dinu, G., Thater, S., and Laue, S. (2012). A comparison of models of word meaning in context. In *Proceedings of the 2012 Conference of the the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 611–615.

Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP 2011*, pages 67–71.

Heylen, K., Peirsman, Y., Geeraerts, D., and Speelman, D. (2008). Modelling Word Similarity. An Evaluation of Automatic Synonymy Extraction Algorithms. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2008)*, pages 3243–3249, Marrakech, Morocco. ELRA.

Heylen, K., Speelman, D., and Geeraerts, D. (2012). Looking at word meaning. an interactive visualization of semantic vector spaces for dutch synsets. In *Proceedings of the EACL 2012 Joint*

 © ATALA

Workshop of LINGVIS & UNCLH, pages 16–24, Avignon, France. Association for Computational Linguistics.

Kievit-Kylar, B. and Jones, M. (2012). Visualizing multiple word similarity measures. *Behavior research methods*.

Lau, J. H., Cook, P., McCarthy, D., Newman, D., and Baldwin, T. (2012). Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics.*, pages 591–601. Association for Computational Linguistics.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Ordelman, R. (2002). Spoken document retrieval for historical video archives - dutch speech recognition in the echo project. Technical report, University of Twente, Parlevink Group.

Peirsman, Y., Geeraerts, D., and Speelman, D. (2010). The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, 16(4) :469–491.

Peirsman, Y., Heylen, K., and Geeraerts, D. (2008). Size matters : tight and loose context definitions in English word space models. In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pages 34–41, Hamburg. ESSLLI.

Rohrdantz, C., Hautli, A., Mayer, T., Butt, M., Plank, F., and Keim, D. A. (2011). Towards Tracking Semantic Change by Visual Analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (Short Papers)*, pages 305–310, Portland, Oregon, USA. Association for Computational Linguistics.

Ruette, T., Geeraerts, D., Peirsman, Y., and Speelman, D. (2012). Aggregating dialectology and typology : linguistic variation in text and speech, within and across languages. In Szmrecsanyi, B. and Wälchli, B., editors, *Linguistic variation in text and speech, within and across languages*. Mouton de Gruyter, Berlin.

Sagi, E., Kaufmann, S., and Clark, B. (2009). Semantic density analysis : Comparing word meaning across time and phonetic space. In *Proceedings of the EACL 2009 Workshop on GEMS : GEometical Models of Natural Language Semantics*, pages 104–111, Athens, Greece.

Schütze, H. (1998). Automatic word sense discrimination. *Comput. Linguist.*, 24(1) :97–123.

Speelman, D. and Geeraerts, D. (2009). The role of concept characteristics in lexical dialectometry. *International Journal of Humanities and Arts Computing*, 2 :221–242.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning : Vector space models of semantics. *CoRR*, abs/1003.1141.

van Noord, G. (2006). At Last Parsing Is Now Operational. In *Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles (TALN06)*, pages 20–42, Leuven, Belgium. Presses universitaires de Louvain.

 © ATALA

Vers un environnement de production et de validation de ressources lexicales sémantiques

Mikaël Morardo Éric Villemonte de La Clergerie
Alpage, INRIA / Université Paris 7,
Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
{mickael.morardo,eric.de_la_clergerie}@inria.fr

RÉSUMÉ
Nous présentons quelques composants d'un environnement pour la production, la visualisation et la validation de ressources lexicales (termes et relations). Au coeur se trouve un composant de construction de réseau lexical entre termes s'appuyant sur l'hypothèse distributionnelle de Harris appliquée aux dépendances syntaxiques produites par l'analyseur FRMG sur gros corpus. Un autre aspect important concerne l'utilisation d'une interface WEB pour la visualisation et la validation collaborative des ressources produites.

ABSTRACT
Towards an environment for the production and the validation of lexical semantic resources

We present some components of a processing chain for the creation, visualization, and validation of lexical resources (terms and relations). At the heart, we find a component for building lexical networks relying on Harris' distributional hypothesis applied on the syntactic dependencies produced by the French parser FRMG on large corpora. Another important aspect concerns the use of a WEB interface for the visualization and collaborative validation of the resulting resources.

MOTS-CLÉS : hypothèse distributionnelle, réseau lexical, extraction terminologique, dépendances syntaxiques, interface WEB collaborative.

KEYWORDS: distributional hypothesis, lexical network, term extraction, syntactic dependencies, collaborative WEB interface.

1 Introduction

Chaque domaine de spécialité possède un ensemble de concepts, souvent exprimés sous forme de termes, qui lui est propre. Pour améliorer les performances de nombreuses tâches (indexation par méta-données, recherche d'information, extraction d'information, aide à la rédaction, etc.), il est important de pouvoir identifier, organiser, et maintenir l'ensemble de ces concepts et de leurs relations, ce rapidement et à coût humain réduit. Nous pensons que cela passe par deux grands axes complémentaires que nous cherchons à concilier dans l'environnement ici présenté.

Le premier axe repose sur l'utilisation de techniques d'acquisition à partir de corpus textuels. Dans notre cas, nous exploitons des corpus analysés syntaxiquement avec l'analyseur FRMG. Les sorties produites permettent l'extraction de termes (multi-mots) et la constitution de réseaux lexicaux

 © ATALA

entre ces termes, en s'appuyant, comme de nombreux prédécesseurs (Bourigault, 2002; Lin & Pantel, 2002), sur l'hypothèse distributionnelle de Haris (Harris, 1968). Cette hypothèse stipule que des mots sémantiquement proches apparaissent dans des contextes similaires et conduit donc à examiner les contextes (syntaxiques dans notre cas) dans lesquelles apparaissent les mots. Elle nous amené à formuler et tester une variation de l'algorithme de regroupement markovien (MCL) (van Dongen, 2000) permettant de rapprocher d'une part les termes, et d'autres part les contextes syntaxiques.

Le deuxième axe guidant le développement de notre environnement concerne la mise en place d'une interface de visualisation et de validation collaborative. Il paraît en effet encore irréaliste de produire automatiquement des ressources parfaites, également difficiles à évaluer, rendant nécessaire une expertise humaine. L'interface permet de facilement parcourir les réseaux de termes, naturellement matérialisés sous forme de graphes, en fournissant des éléments d'explication (dont des exemples) et des fonctionnalités d'édition pour des validations plus rapides.

La section 2 présente le matériel de départ pour nos expériences, utilisé en particulier pour de l'extraction de termes multi-mots (section 3) et la constitution de réseaux lexicaux traduisant la proximité sémantique (section 4). Des éléments d'analyse des résultats sont fournis en section 5 avant de présenter section 6 les grandes lignes de notre interface de visualisation et de validation, implantée dans le cadre de la plateforme Libellex qui ouvre la voie à des investigations plus poussées des ressources produites.

2 Corpus annotés syntaxiquement

Nous avons mené plusieurs expériences d'acquisition sur divers corpus, spécialisés ou non. Ainsi, `All` (740 Millions de mots) est un corpus composite et généraliste [1] qui regroupe, entre autres, Wikipedia (180Mmots), Wikisource (littéraire, 64Mmots), et des dépêches AFP (journalistique, 248Mmots). Nous exploitons actuellement un corpus juridique (470Mmots) découpé en plusieurs sous-domaines (fiscal, affaires, social, civil et comptable). Nous traitons aussi des corpus beaucoup plus petits, comme un de description botanique (flores) ou un de maintenance automobile. Ces corpus sont analysés syntaxiquement avec l'analyseur du français à large couverture FRMG (Villemonte de la Clergerie, 2010), intégré dans la chaîne de traitement linguistique ALPAGE.

Pour nos expériences, les résultats d'analyse sont fournis au format Passage (Vilnat *et al.*, 2010), moins riche que le format DepXML nativement fourni par FRMG, mais produit par au moins une dizaine d'analyseurs syntaxiques pour le français. Le format Passage s'appuie sur 6 types de *chunks* (GN, GP, NV, PV, GA [adjectifs], GR [adverbes]) et 14 types de dépendances entre formes et/ou *chunks*, que l'on peut facilement ramener systématiquement à des dépendances entre formes. À partir des résultats d'analyses, certains éléments récurrents sont collectés et comptés, en s'appuyant sur une adaptation de l'algorithme MapReduce (Dean & Ghemawat, 2004).

1. Le choix de ce corpus général s'explique par des raisons historiques et également pour s'intéresser à un cas plus difficile que celui des domaines spécialisés, qui restent néanmoins le champ d'action naturel pour ce genre de travaux.

 © ATALA

3 Extraction terminologique

Reprenant des idées bien connues (Pazienza *et al.*, 2005), l'extraction de termes candidats multi-mots s'appuie en premier lieu sur la collecte des séquences de *chunks* correspondant (approximativement) au motif `(GN)(GR*GA|GP|PV)+`[2], les composants étant liés syntaxiquement (essentiellement par des relations MOD-N pour *modifieur de nom*). Les candidats sont classés en fonction de multiples critères comme leur fréquence, leur cohésion interne (calculée via une mesure d'information mutuelle point à point), leur autonomie, et leur diversité d'utilisation. L'autonomie exploite les dépendances pour vérifier qu'un terme peut jouer des rôles actifs (comme sujet et objet) et n'est pas systématiquement modifié (par des GP par exemple). La diversité part du constat que certaines séquences sont parfois des collocations apparaissant dans des phrases clichées, souvent largement identiques, ce que nous pénalisons. Les variants sont regroupés en fonction des lemmes sous-jacents, et certains candidats sont rejetés à cause d'une trop grande variabilité (provenant par exemple d'un lemme `_NUMBER` ou `_DATE` correspondant à une entité nommée).

Avec un filtrage minimum, nous obtenons de l'ordre de 100K termes sur les sous-corpus AFP et Wikipedia et 50K sur la partie *fiscale* du corpus juridique (145M mots). Les termes extraits viennent avec un ensemble de phrases illustratives et des informations statistiques. La figure 3 montre quelques exemples de termes issus de la partie `affaires`.

4 Regroupement sémantique

Le regroupement sémantique des mots simples et des termes extraits précédemment s'appuie sur la collecte des dépendances syntaxiques, retournées sous forme de triplets (gouverneur, relation, gouverné) comme par exemple le triplet (*se_asseoir*, *sur*, *chaise*).

Nous ignorons certaines relations Passage (comme la juxtaposition) et appliquons certaines modifications pour essayer de nous rapprocher du niveau sémantique. Ainsi, sans être exhaustif :

- les passifs sont réécrits pour que le sujet de surface devienne un objet, et le complément introduit par *par* un sujet profond ;
- la relation ATB-SO (attribut du sujet ou de l'objet) qui lie le verbe à l'attribut relie l'attribut directement au sujet ou à l'objet ;
- la relation SUJ-V pour Passage s'attache sur le premier verbe, qui peut être l'auxiliaire : nous retrouvons alors le vrai verbe ;
- les relations mettant en jeu un coordonnant sont redistribuées sur les éléments coordonnés ;
- les modifieurs prépositionnels (MOD-N ou CPL-V) donnent un triplet précisant la préposition ;
- les prépositions sont postfixées par = quand elles introduisent un groupe nominal sans article ;
- des triplets supplémentaires sont ajoutés pour certains cas d'ambiguïtés d'attachement prépositionnels. Ainsi, dans une expression comme *journal de bord du capitaine*, la désambiguisation propose le triplet (*bord*,*de*,*capitaine*) mais nous ajoutons le triplet indirect (*journal*,*de**,*capitaine*) avec la marque *. La motivation est que les erreurs d'attachement induites par ces triplets indirects vont se distribuer sur de nombreux gouverneurs mais que les bons attachements (sur la tête d'un terme multi-mots) vont s'accumuler. La marque permet de filtrer

2. le `GN` initial peut aussi être un `GP` sans sa préposition.

© ATALA

ces triplets, et éventuellement de leur appliquer une confiance plus faible.
– dans la lignée du point précédent, des triplets sont aussi ajoutés quand le gouverneur ou le gouverné est tête d'un terme multi-mots issu de la phase d'extraction. L'exemple précédent ajouterait donc un triplet entre *journal de bord* (si détecté comme un terme) et *capitaine*. Dans `affaire` pour le terme *procédure collective*, on trouve ainsi 6462 occurrences de (*ouverture*,*de*,*procédure collective*), ou encore 189 occurrences pour (*qualité*,*de*,*président du conseil*).

La table 1 montre quelques exemples fréquents de triplets de dépendances, issus de `All`, impliquant le mot *chaise*. On constate la fréquence forte de la coordination ainsi que la forte représentation de triplets traduisant des termes (*chaise longue*, *chaise de poste*, *chaise musicale*, *chaise à porteur*, ...). Les triplets que l'on considère comme représentatifs de *chaise*, comme *s'asseoir sur*, bien que présents, sont loin d'être les seuls à forte fréquence.

gouverneur	relation	gouverné	fréq.	gouverneur	relation	gouverné	fréq.
chaise_nc	et	table_nc	**235**	prendre_v	cod	chaise_nc	**87**
asseoir_v	sur	chaise_nc	**227**	chaise_nc	modifieur	électrique_adj	**82**
chaise_nc	modifieur	long_adj	**168**	chaise_nc	modifieur	vide_adj	**80**
chaise_nc	de=	poste_nc	**115**	chaise_nc	à=	porteur_nc	**80**
tomber_v	sur	chaise_nc	**103**	dossier_nc	de	chaise_nc	**78**
chaise_nc	modifieur	musical_adj	**102**	avoir_v	cod	chaise_nc	**71**
se_asseoir_v	sur	chaise_nc	**93**	table_nc	et	chaise_nc	**62**

TABLE 1 – Quelques exemples de triplets de dépendance impliquant *chaise*.

Un triplet comme (*se_asseoir*,*sur*,*chaise*) permet d'associer le contexte <*se_asseoir sur* •> à *chaise* et le contexte <• *sur chaise*> à *se_asseoir*. Chaque mot w_i se voit ainsi associé un vecteur $[c_1 : u_{1i}, \cdots]$ de contextes c_a (avec leurs décomptes u_{ai}) qui vont servir à calculer la similarité. Comme ces vecteurs peuvent avoir un très grand nombre de composants, nous éliminons les contextes non lexicalisés par un mot plein (faisant par exemple référence à un pronom, comme <• *cln sujet*>), les contextes associés à un unique mot, et ceux à trop forte fréquence. De plus, les arguments phrastiques (correspondant à des complétives ou des infinitives) comme <• *cod manger*> sont réécrits en <• *cod* `*sentence*`>, considérant que, en première approximation, les verbes ainsi enchâssés ne sont pas réellement pertinents pour le regroupement de mots. Finalement, les mots à très faible fréquence (< 10 par défaut) sont éliminés.

Les paires (*mot* w_i,*contexte* c_a) fournissent les arcs d'un graphe biparti sur lequel nous adaptons l'algorithme de *clustering* markovien (MCL) (van Dongen, 2000). Celui-ci considère que des mots sémantiquement proches doivent être liés par un ensemble dense de chemins courts dans un graphe connectant les mots. L'algorithme originel s'appuie sur une matrice de contingence *mot-mot* et applique un opérateur d'*inflation/normalisation* tendant à renforcer la part relative des chemins courts et à diminuer celle des chemins longs. Dans notre cas, nous souhaitons préserver la dualité entre mots et contextes avec notre graphe biparti liant des mots w_i à des contextes c_a par des arcs orientés de poids $w_i \bullet c_a$ de w_i vers c_a et de poids $c_a \bullet w_i$ de c_a vers w_i. L'objectif est alors de connecter les mots par des poids $w_i \bullet w_j$ dénotant leur similarité et, similairement, de connecter les contextes par des poids $c_a \bullet c_b$, comme illustré par la figure 1.

Les formules (1) traduisent ces idées, en forme développée à gauche et en forme matricielle compacte à droite avec $W = (w_i \bullet w_j)$ la matrice de similarité sur les mots et $C = (c_i \bullet c_j)$ celle sur les contextes.

 © ATALA

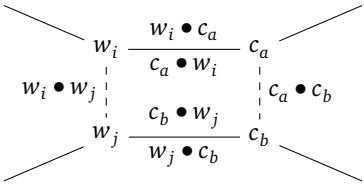


FIGURE 1 – Graphe termes – contextes pour l'algorithme de regroupement

$$\begin{cases} w_i \bullet w_j = \dfrac{1}{Z_i}\left(\displaystyle\sum_{a,b}(w_i \bullet c_a)(c_a \bullet c_b)(w_j \bullet c_b)\right)^{\alpha} & W = \Gamma_\alpha(F^t C F) \\ c_a \bullet c_b = \dfrac{1}{Z_a}\left(\displaystyle\sum_{i,j}(c_a \bullet w_i)(w_i \bullet w_j)(c_b \bullet w_j)\right)^{\alpha} & C = \Gamma_\alpha(G^t W G) \end{cases} \tag{1}$$

où les Z_i et Z_a dénotent les facteurs de normalisation donnés par

$$\begin{cases} Z_i = \displaystyle\sum_{j}\left(\sum_{ab}(w_i \bullet c_a)(c_a \bullet c_b)(w_j \bullet c_b)\right)^{\alpha} \\ Z_a = \displaystyle\sum_{b}\left(\sum_{i,j}(c_a \bullet w_i)(w_i \bullet w_j)(c_b \bullet w_j)\right)^{\alpha} \end{cases} \tag{2}$$

La forme développée stipule que la similarité $w_i \bullet w_j$ se calcule en sommant les contributions des divers chemins allant de w_i à w_j en passant par des contextes c_a et c_b, ce qui, récursivement, résume la contribution $c_a \bullet c_b$ des chemins allant de c_a à c_b. Le facteur d'inflation $\alpha \geq 1$, appliqué à ces contributions, augmente la part relative (après normalisation) des chemins forts (normalement les plus courts passant par des arcs de poids élevé). Par défaut, nous fixons $\alpha = 2$ mais ce taux est en fait modulé par le taux de connectivité des parties du discours. [3]

Sous la forme matricielle compacte, l'**opérateur d'inflation** Γ_α combine inflation et normalisation tandis que la matrice F (et sa transposée F^t) regroupe les poids $(w_i \bullet c_a)$ du contexte c_a pour le mot w_i (similaire pour G sur les contextes). Comme nous considérons qu'un contexte c_a utilisé avec de nombreux mots est moins intéressant qu'un contexte plus spécifique de seulement quelques mots (mais pas d'un seul mot !), nous pondérons les poids $w_i \bullet c_a$ selon la formule suivante (similairement pour $c_a \bullet w_i$ dans G) :

$$w_i \bullet c_a = \frac{ln(u_{ai}) * \eta_a}{\sum_b ln(u_{bi}) * \eta_b} \quad \text{avec} \quad \eta_a = ln\left(\frac{\#\text{mots distincts}}{\sqrt{|\{w_j | u_{aj} > 0\}|}}\right) \tag{3}$$

Les deux formules de l'équation (1) sont mutuellement récursives et le calcul est approché à l'aide d'un algorithme itératif de point fixe en démarrant avec des matrices initiales $W^{(0)}$ et $C^{(0)}$. Nous obtenons ainsi une version de base de l'algorithme.

3. Nos expériences ont en effet montré que certaines parties du discours (comme les noms communs) sont moins connectables que d'autres (comme les adverbes).

© ATALA

Cependant, nous étendons ce modèle de base, pour pouvoir explorer diverses intuitions linguistiques et utiliser des informations externes. [4] Ainsi, comme les contextes sont structurés et enchâssent en fait un mot en leur sein, nous considérons que les contextes $r.w_i$ et $r.w_j$ construits à partir de 2 mots similaires w_i et w_j avec une même relation r (par exemple `sujet`) ont des chances d'être également similaires. Réciproquement, si $r.w_i$ et $r.w_j$ sont similaires, alors w_i et w_j ont des chances d'être similaires. Ceci peut se formaliser par l'équation (4), en utilisant un opérateur τ de transfert des contextes vers les mots, et, réciproquement, un opérateur ρ de transfert des mots vers les contextes, avec, pour chaque relation r, un coefficient de transfert β_r (à 0,2 par défaut) et une matrice de transfert $T_r = (\langle w_i \bullet c_a \rangle_r)_{ia}$ avec $\langle w_i \bullet c_a \rangle_r = 1$ si $c_a = r.w_i$ et 0 autrement.

$$\begin{cases} W = \Gamma_\alpha(F^t C F + \tau(C)) \text{ avec } \tau(C) = \sum_r \beta_r T_r^t C T_r \\ C = \Gamma_\alpha(G^t W G + \rho(W)) \text{ avec } \rho(W) = \sum_r \beta_r T_r W T_r^t \end{cases} \tag{4}$$

Par ailleurs, l'équation (1) n'assure pas l'auto-similarité d'un mot avec lui-même, même partant de matrices initiales avec $w_i \bullet w_i = 1$. Cependant, il est possible d'assurer cette auto-similarité en réinjectant à chaque itération un bonus multiplicatif dans le calcul de $w_i \bullet w_i$. De manière générale, ceci s'exprime par la formule (5), avec utilisation de matrices $L = (l_{ij})$ et $K = (k_{ab})$ de bonus de similarité, pour lesquelles $l_{ij} = (i == j)$ et $k_{ab} = (a == b)$. Dans cette formule, I désigne la matrice identité et $\circ$ le produit point à point de Hadamard avec $(A \circ B)_{ij} = A_{ij}.B_{ij}$. Après application des bonus, les matrices sont renormalisées, mais sans inflation (Γ_1).

$$\begin{cases} W = \Gamma_1((I + L) \circ \Gamma_\alpha(F^t C F + \tau(C))) \\ C = \Gamma_1((I + K) \circ \Gamma_\alpha(G^t W G + \rho(W))) \end{cases} \tag{5}$$

Cette formulation générique permet d'étendre la notion de bonus à d'autres cas que l'auto-similarité. Ainsi, nous attribuons de tels bonus, plus ou moins importants, dans les cas suivants :

- en présence de coordinations entre w_i et w_j, dépendant du nombre d'occurrences ;
- en cas de typographie proche, par exemple pour des variations portant sur des diacritiques ;
- en cas d'équivalence modulo certains préfixes et suffixes, pouvant traduire une relation de morphologie dérivationnelle ;
- en cas d'utilisation d'une ressource germe fournissant des indications de similarité, par exemple une ressource de type wordnet.

Ces matrices sont également utilisées pour insérer des malus, par exemple entre mots de catégories syntaxiques incompatibles. Nous pénalisons aussi les rapprochements entre mots trop disparates en terme de fréquence ou de diversité de contextes. En effet, nous considérons qu'un mot w_1 très fréquent et productif ne doit pas être rapproché directement d'un mot w_2 rare et spécifique, car il existe probablement un mot intermédiaire w_3 en terme de fréquence, proche de w_1 et w_2. L'espoir est de voir ainsi émerger une hiérarchisation entre les mots.

Pour une paire de mots (w_i, w_j), un des intérêts de la formule (1), toujours présent dans ses variantes, est de pouvoir estimer l'importance $\langle w_i \bullet w_j, c_a \rangle$ d'un contexte c_a intervenant dans le

4. ce qui est une des raisons motivant le développement d'un nouvel outil de regroupement de mots.

© ATALA

rapprochement de w_i et w_j avec

$$\left\langle w_i \bullet w_j, c_a \right\rangle = \sum_b (w_i \bullet c_a)(c_a \bullet c_b)(w_j \bullet c_b) \tag{6}$$

Nous ne conservons alors que les contextes les plus significatifs expliquant (partiellement) la similarité $w_i \bullet w_j$. Il devient aussi possible d'ordonner ces contextes et de rechercher une corrélation entre le classement $\mathscr{C}_{ij}$ ainsi obtenu et le classement originel $\mathscr{D}_i$ des contextes pour w_i. Si $\mathscr{C}_{ij}$ est largement inclus dans le bon ordre dans $\mathscr{D}_i$ avec des poids assez similaires, alors on peut supposer que le rapprochement de w_j avec w_i ne remet pas en cause l'importance relative des contextes pour w_i. Nous mesurons cette similarité entre classements avec la formule (7), où rank_{ai} dénotant le rang de c_a dans $\mathscr{D}_i$ et rank_{aij} son rang dans $\mathscr{C}_{ij}$.

$$\Delta_{ij} = \sum_{c_a, w_i \bullet c_a > 0} \left(1 + \left|1 - \frac{\text{rank}_{aij}}{\text{rank}_{ai}}\right|\right)^{-1} \left(1 + \left|1 - \frac{\left\langle w_i \bullet w_j, c_a \right\rangle}{w_i \bullet c_a}\right|\right)^{-1} \tag{7}$$

Pour des raisons d'efficacité, à chaque itération, nous ne gardons, pour un mot w_i, que les mots w_j dont le niveau de similarité $w_i \bullet w_j$, modulé avec Δ_{ij}, dépasse un certain seuil. Par défaut, nous requérons que $\Delta_{ij} w_i \bullet w_j > 0.01$.

Toujours pour des raisons d'efficacité, nous ne conservons dans $\mathscr{D}_i$ pour w_i que les contextes c_a dont l'importance relative $w_i \bullet c_a$ dépasse un certain seuil ($0,2 \max_a(w_i \bullet c_a)$ par défaut). Dans les cas où les parts relatives sont très uniformes, nous ne conservons que les 50 premiers. Le mauvais aspect de ce choix est de focaliser la recherche de similarité entre mots sur ces au plus 50 contextes (qui diffèrent pour chaque mot), ce qui induit certains effets pernicieux dans le cas de mots très productifs en tant que constructeurs de termes. On a alors des contextes forts découlant des termes en question (comme <• *de terre*> pour *pomme*) qui ne sont cependant pas vraiment significatifs pour le mot en question (comme *pomme*). Pour remédier à cette situation, avant d'éliminer des contextes pour w_i, nous construisons un vecteur ri_i indiquant la présence (1) ou l'absence (0) dans les contextes de w_i de 1000 contextes initialement choisis aléatoirement, selon les principes du *Random Indexing* (Kanerva *et al.*, 2000). Ces vecteurs offrent une vue plus large sur les contextes des mots. En cas de rapprochement de w_i et w_j, un bonus/malus $b_{ij} = cos(r_i, r_j) - 1$ est ajouté dans la matrice L.

5 Analyse des résultats

En sortie de l'algorithme, nous obtenons une liste ordonnée de paires non symétriques de mots, avec, pour chaque paire (w_i, w_j), des informations sur la force du rapprochement et une liste de contextes significatifs c_a ordonnés selon $\left\langle w_i \bullet w_j, c_a \right\rangle$. Pour le corpus cpl [5], nous obtenons ainsi 19 960 mots w_i et 51 980 paires (w_i, w_j), ce qui représente déjà un nombre important de connexions. En symétrisant ce graphe orienté, nous obtenons un graphe non orienté de 20 922 nœuds et 47 065 arêtes. Pour le corpus `affaire`, nous obtenons 10 223 noeuds et 13 584 arcs. [6]

5. pour *Corpus Passage Long* (495Mmots), correspondant à `all` moins la sous-partie AFP.
6. le nombre d'arêtes est a priori quadratique en le nombre de noeuds, avec ici des ratios a^2/n de 0,010 et 0,011.

© ATALA

La table 2 donne des indications sur la connectivité en fonction de la catégorie syntaxique pour divers corpus. On constate que la majorité des mots connectés sont en fait des entités nommées, mais que cela ne représente qu'une infime partie de celles-ci (dû au fait que l'immense majorité des entités ont moins de 10 occurrences). Dans une moindre mesure, la situation est similaire pour les noms communs. Par contre, les adverbes se connectent très facilement. En conséquence, l'algorithme d'acquisition a été légèrement modifié pour tenter de corriger (mais pas totalement) ces distorsions en modulant le facteur d'inflation en fonction de la catégorie syntaxique, avec par exemple une inflation de $\alpha + 4$ pour les adverbes et de $\alpha - 1$ pour les noms communs. On note que les termes sont légèrement moins connectables que les noms communs, ce qui pourrait s'expliquer par le fait qu'un terme a vocation à éliminer les formulations synonymes.

corpus	**partie du discours**	**#mots**	**#conservés**	**%c/m**
cpl	np	1 779 848	8 749	0,5
	nc	35 417	5 782	16,3
	v	11 224	2 480	22,1
	adj	10 198	3 108	30,5
	adv	2 693	776	28,8
fiscal	terme	50 479	4 981	9,9
	nc	18 760	1 976	10,5
	v	4 783	593	12,4
affaires	terme	65 138	5 142	7,9
	nc	23 506	2 095	8,9

TABLE 2 – Distribution de la connectivité sur quelques corpus et parties du discours.

En examinant la liste des paires, on constate que l'algorithme regroupe correctement certaines paires de mots correspondant à des erreurs typographiques comme *boîte* et *boite*. On retrouve également des regroupements correspondant à des variations orthographiques, comme *clé* et *clef*, *lis* et *lys*, ou *audio-visuel* et *audiovisuel*. Noté dans les premières versions de l'algorithme, ce phénomène a ensuite été renforcé par l'octroi d'un faible bonus en cas de proche distance d'édition (en particulier sur les diacritiques).

D'autre part, on observe une forte proportion de rapprochements s'appuyant partiellement sur le bonus de coordination (25,7% pour **cpl**). Dans nombre de cas, ce bonus ne vient cependant que conforter une décision motivée par des contextes, comme par exemple pour *chaise* et *divan*.

TULIP (Auber, 2003), un outil de visualisation de très grands graphes, permet d'avoir une vue plus globale du réseau. La figure 2 montre ainsi un fragment du graphe pour `all` reliant des parties du corps humain (principalement les os et muscles). TULIP permet déjà d'avoir une impression subjective mais utile de la qualité du réseau lexical. Il fournit aussi des informations précieuses sur la topologie globale et locale du réseau.

En particulier, on observe rapidement la présence de *buissons* de mots fortement interconnectés. En s'appuyant sur quelques considérations topologiques et des mesures standards dans les graphes [7], nous avons pu ainsi repérer ces buissons et en extraire environ 4000 classes (pour **cpl**), avec par exemple :

7. l'algorithme s'appuie sur la forte proportion de connections entre les fils N_i d'un noeud N (indiquant son appartenance à un buisson) et sur la diffusion de couleurs entre noeuds d'un même buisson pour en trouver les limites.

 © ATALA

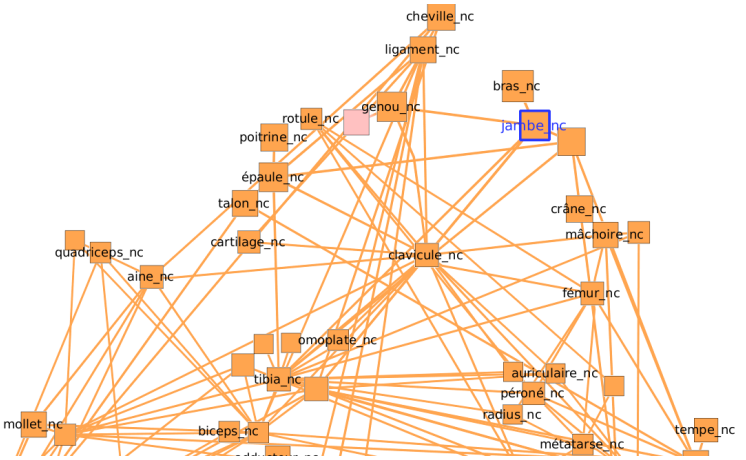


FIGURE 2 – Fragment de réseau, autour de *jambe*, visualisé sous le logiciel Tulip

<79> *sulky malinois fox-terrier setter cocker colley chiot fox labrador ratier griffon caniche teckel épagneul*

<80> *arrière-garde canonnier cavalerie carabinier tirailleur hussard panzer voltigeur blindé grenadier cuirassier avant-garde zouave lancier*

<83> *pneumonie paludisme diphtérie pneumopathie variole dysenterie malaria botulisme poliomyélite septicémie varicelle polio rougeole méningite*

Néanmoins, la vue avec TULIP reste subjective et ne dispense pas de tenter de mener des évaluations plus objectives. Nous en avons menées deux différentes, mais, pour des raisons de place, ne présentons ici que celle inspirée du test TOEFL [8] (Turney, 2002), en l'appliquant sur des ressources de type WordNet. Plus précisément, nous utilisons : [9]

– le WordNet du Français (**fwn**) (Jacquin *et al.*, 2007), qui ne couvre que les catégories nominales et verbales. Il comporte 22 121 synsets ;
– WOLF, un wordnet du français, aligné sur les synsets du wordnet anglais, automatiquement construit à partir de textes multilingues alignés et couvrant plus de catégories syntaxiques en ajoutant les adjectifs et les adverbes. Il comporte 32 351 synsets non vides regroupant 38 001 lexèmes distincts (Sagot & Fišer, 2008) ;

Un tel test TOEFL comprend une liste de questions, chaque question spécifiant un terme candidat et une liste de quatre termes réponses dont seulement un est proche sémantiquement du candidat. Le but du jeu est donc de retrouver la bonne réponse pour chaque question, avec une probabilité de réussite de 25% pour un programme répondant au hasard. N'ayant pas de test TOEFL sous la main pour le français, nous les engendrons à partir des wordnets, en choisissant le candidat et, aléatoirement, la bonne réponse (dans le même *synset*) ainsi que 3 autres termes leurres (dans d'autres *synsets*), comme illustré ci-dessous :

toutefois	*néanmoins*	complètement	progressivement	sensiblement
exploit	*prouesse*	offset	plie	bit

8. *Test of English as a Foreign Language*
9. Nous avons aussi mené des évaluations sur d'autres versions de WordNet qui nous omettons ici faute de place.

© ATALA

Néanmoins, on observe un faible recouvrement entre les vocabulaires des wordnets et du réseau. Nous filtrons donc les éléments du test, en demandant que les termes aient au moins 100 occurrences dans notre corpus, et ensuite pour une instance de réseau à évaluer, en demandant que les termes soient présents dans le réseau. Pour chaque réseau et wordnet, nous n'évaluons donc pas exactement sur le même jeu de test. En augmentant la taille des jeux, nous espérons néanmoins obtenir des mesures représentatives. Pour répondre aux tests, nous choisissons la réponse la plus proche de la question grâce à un chemin dans le réseau, et utilisons un tirage aléatoire en l'absence de chemins. Les résultats sont donnés par la table 3 pour plusieurs corpus sur les deux wordnets. Comme *baseline*, nous faisons aussi passer le test au wordnet WOLF, vu comme un réseau sémantique.

	fwn		**wolf**	
corpus	**%ok**	**#tests**	**%ok**	**#tests**
all	51,5	4 121	42,1	7 674
fiscal	46,1	104	37,0	493
affaires	35,1	248	43,2	1 055
social	39,4	274	37,7	1 345
wolf	64,5	1 076		

TABLE 3 – Évaluations de type Toefl.

On constate que les résultats sont en dessous de la référence fournie par le réseau wolf, ce qui n'est pas totalement surprenant, vu que WOLF a été certes automatiquement construit mais en s'appuyant totalement sur la structure du wordnet anglais. Les résultats sont néanmoins largement supérieurs à la *baseline* hasard de 25%. On note que les résultats sont meilleurs sur le French WordNet, mais avec une plus faible couverture en terme de vocabulaire (car FWN ne couvre que les catégories nominales et verbales). Pour les corpus spécialisés, les résultats sont, sans surprise, moins significatifs (moindre recouvrement du vocabulaire) et moins bons (en partie à cause d'une référence non spécialisée).

La table 4 fournit des résultats d'évaluation plus fins au niveau des catégories syntaxiques, la colonne **bons** (resp. **mauvais**) donnant le ratio de bonnes (resp. mauvaises) réponses obtenues en trouvant un chemin dans le réseau, tandis que **manquants** indique le ratio de cas où le choix de la réponse (bonne ou mauvaise) s'est faite par tirage au sort. On constate que, quand des chemins existent pour un `nc`, la réponse est presque toujours bonne (très forte précision) mais qu'on souffre de la faible connectivité des noms communs (faible rappel). À l'autre extrême, les adverbes sont facilement connectables, mais les réponses sont majoritairement fausses. Les verbes occupent une position intermédiaire.

wordnet	**pos**	**#tests**	**%bons**	**%faux**	**%manquants**	$\%b/(b+f)$
wolf	v	3 876	35,5	30,9	33,6	53,4
	nc	1 078	33,5	2,1	64,4	94,0
	adj	2 085	22,3	11,3	66,4	66,3
	adv	1 533	36,9	41,9	21,7	46,8

TABLE 4 – Tests Toefl par catégories syntaxiques (sur `cpl`).

© ATALA

6 Visualiser et valider avec Libellex

La situation à ce stade n'est pas totalement satisfaisante. Les vues fournies par un outil comme TULIP montrent la qualité de larges portions du réseau, tout en permettant de détecter, sans possibilité de corrections, un certain nombre d'erreurs. Les évaluations confirment que les réseaux ne sont pas parfaits mais démontrent aussi la difficulté de trouver des ressources de référence (il suffit de noter le manque d'accord entre French WordNet et Wolf). Le passage à des domaines spécialisés, sans référence existante et avec des réseaux de taille conséquente, pose aussi problème.

Dans le cadre d'une collaboration avec la société Lingua & Machina (L&M) pour étendre leur plateforme LIBELLEX, nous avons développé une interface Web pour parcourir, visualiser et valider des ressources lexicales, incluant termes et relations. Cette interface a pour objectif de répondre à plusieurs points essentiels :

- explorer facilement les résultats pour permettre aussi bien la mise en évidence de phénomènes dans la langue que d'erreurs dans les algorithmes.
- évaluer les résultats à la main tout en offrant une ergonomie facilitant la validation (les marquer comme acceptable ou non-acceptable avec une granularité variable).
- partager et échanger entre les utilisateurs sur des points ou des résultats ambiguës, difficiles à interpréter ou évaluer.
- conserver une trace des actions effectuées par un humain.

L'interface affiche une liste de termes (avec des options de tri et de recherche) avec les différentes informations associées au terme sélectionné (explications et phrases d'exemple). L'interface essaie d'être intuitive pour la navigation et les fonctions de validation (par exemple pour accepter ou rejeter un terme, changer le représentant vedette, enlever ou ajouter un variant, etc., pour ne parler que des termes) et doit faciliter le travail pour un validateur expérimenté ou débutant.

Le volume de données étant relativement grand (plus de 100K termes candidats pour la plus grosse ressource traitée dernièrement), travailler rapidement devient un impératif si l'on souhaite bénéficier dans un temps acceptable de données validées. L'interface offre la possibilité de communiquer entre utilisateurs pour partager leur expérience sur des points délicats (validité d'un terme par exemple) à l'aide d'un système de commentaires. En plus de cela, un utilisateur peut alerter sur des termes qu'il considère comme nécessitant une attention particulière.

Pour représenter les relations entre termes, la facilité de lecture qu'offrent les graphes a retenu notre attention, comme illustré par la figure 3. D'un coup d'oeil il est facile de repérer des anomalies comme des relations qui semblent dépourvues de sens ou des noeuds aux labels étranges. C'est pourquoi nous avons adopté un système de visualisation par graphe en s'appuyant sur la bibliothèque javascript `d3.js`. Pour ne pas surcharger l'utilisateur, seuls les voisins à deux générations du terme sélectionné sont visualisés, mais un click sur un noeud permet de recentrer la vue sur celui-ci. La taille des noeuds de seconde génération indique leur degré de connectivité. Il est également possible de rechercher les plus courts chemins entre deux termes.

Nous avons essayé de maintenir un graphe cohérent à l'affichage. Certaines parties d'un graphe peuvent contenir des sous-ensembles significatifs (des *clusters*) que souhaitons mettre en avant pour qu'ils soient identifiables au premier coup d'oeil. Ainsi, l'algorithme de placement des noeuds, qui s'appuie sur l'utilisation de forces attractives (intra-cluster) et répulsives (inter-cluster) permet de mieux séparer visuellement les divers sens qu'un terme possède (sous réserve

 © ATALA

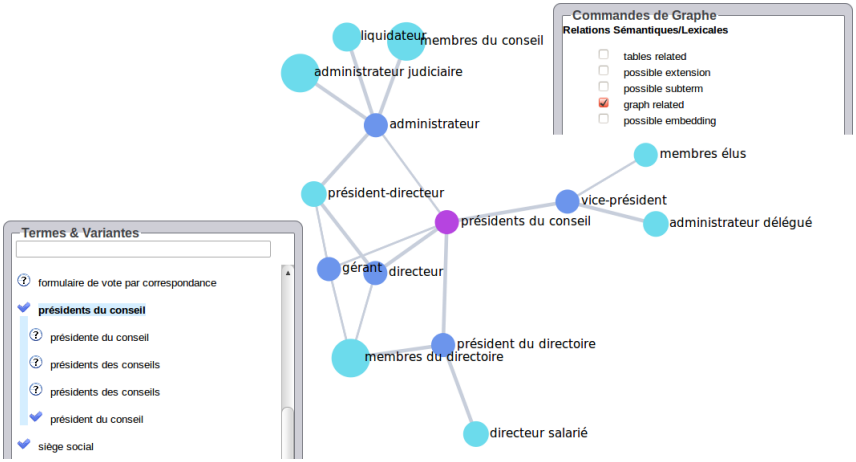


FIGURE 3 – Visualisation d'une partie du réseau sous Libellex (corpus `affaires`)

que l'algorithme d'extraction les ait trouvés). Les fonctionnalités d'édition permettent d'ajouter ou supprimer temporairement des relations (et prochainement de les valider).

Pour aider à prendre une décision lorsque l'on étudie le lien entre deux noeuds (ou plus), il est possible de consulter une matrice affichant, pour chaque terme sélectionné, les contextes les plus représentatifs ayant servi au rapprochement avec leur poids (indiqué par une jauge de couleur). Cette matrice fournit une vue synthétique et efficace pour vérifier la pertinence d'un lien entre deux noeuds. L'affichage de phrases d'exemples pour illustrer les contextes sert à lever toute ambiguïté. Il est à noter que cet outil a démontré son utilité dans de nombreux cas où l'intuition linguistique première d'un utilisateur était de considérer comme faux un lien donné.

Il est à noter que les matrices de contextes donnent de bons indices par rapport à la polysémie d'un terme. Ainsi, les mots *char*, *charrette* et *chariot* partageant les contextes <• *à bœuf*>, <• *modifieur atteler*>, <*promenade en* •> et d'autre part *char* et *tank* partageant les contextes <• *de combat*>, <*régiment de* •>, <• *modifieur lourd*>.

Nous pouvons donc vérifier, à différents degrés, si les données produites sont bonnes et, plus généralement, s'il n'y a pas d'erreurs dans les outils d'acquisition. En parcourant de manière locale les données, il devient plus simple de redonner du sens à des résultats statistiques mais il est aussi possible d'observer des phénomènes de langue plus difficilement identifiables avec d'autres formes de visualisation. Dans l'exemple de la figure 4, nous avons conclu que les contextes servant à rapprocher les termes sont, dans l'ensemble, liés au journalisme sportif. Il s'avère que la nature du corpus oriente les relations entre les termes. Il existe sûrement divers liens sémantiques entre *orteil* et *poignet*, mais, dans le corpus `all`, leur proximité est due à leur capacité à être le siège d'une blessure (*cassure*, *fracture*, etc.) pour des sportifs.

Sur un autre plan, un peu plus en marge, nous avons constaté l'intérêt que suscitent des graphes de mots dynamiques et interactifs au sein d'une population variée d'utilisateurs. C'est un phénomène intéressant à noter car démontrant un plaisir d'utilisation de la plateforme dans l'exploration des données : il y a un aspect ludique involontaire qui se dégage et qui mérite notre

 © ATALA

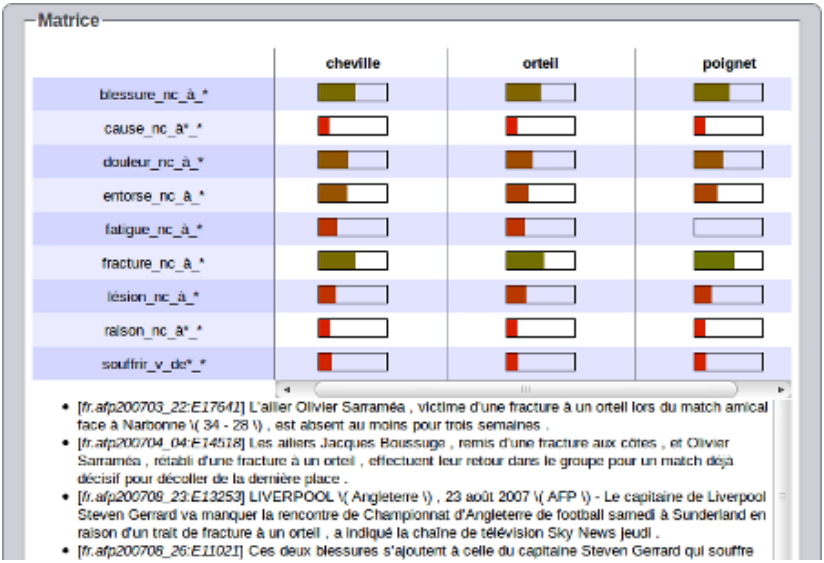


FIGURE 4 – Visualisation d'une matrice de contextes avec exemples

attention à l'avenir. En effet, éviter que l'utilisateur ne s'ennuie devant une tâche répétitive est un problème que nous considérons en permanence lors du développement. Nous avons donc une plateforme qui permet à la fois une exploration et une validation des données avec facilité et avec rapidité en plus d'un aspect collaboratif (à travers notamment le système de commentaires et de mise en avant d'un terme).

7 Conclusion

Nous avons présenté un environnement pour la production et la validation de réseaux lexicaux sur des termes simples ou multi-mots. Le coeur de cet environnement est un algorithme de regroupement de mots, flexible et capable de prendre en compte de nombreux paramètres, dont certains s'appuyant sur des intuitions linguistiques, comme par exemple la structure des contextes syntaxiques, les coordinations, des éléments de morphologie dérivationnelle et la distance d'édition.

Par ailleurs, bien que l'algorithme d'acquisition ait été conçu en prenant en compte la dualité des mots et des contextes, nous n'avons pas encore pleinement exploité ces derniers, même si l'interface de visualisation permet déjà de basculer sur une vue de contexte.

L'interface de visualisation et de validation collaborative, intégrée au sein de la plateforme Web LIBELLEX, nous semble un composant de plus en plus utile pour pouvoir mener des tâches d'évaluation plus précises (pour nous) et pour produire rapidement des ressources de qualité sur un domaine spécialisé (pour les utilisateurs). Comme l'expertise humaine est une chose précieuse, l'un des enjeux est de s'assurer que le travail d'un humain n'est jamais vain. Il doit d'abord être

 © ATALA

conservé et non-altéré. Ensuite, nous devons l'utiliser pour des tâches d'évaluation et pour fournir du retour pour améliorer la chaîne de traitement (aux niveaux linguistique et acquisition). Enfin, les données validées doivent être exploitées dans les outils d'acquisition, par exemple sous forme de ressource germe dans la phase de regroupement de mots, ou pour entraîner des classifieurs pour mieux fixer les paramètres des algorithmes d'acquisition. On voit ainsi se mettre en place un cercle vertueux.

Il est à noter que nombre d'éléments de l'environnement sont remplaçables, que ce soit la chaîne linguistique, le système d'extraction de termes, ou celui de regroupement de mots. Le composant qui reste le plus indispensable est peut-être au final l'interface de visualisation et de validation.

Cette interface est accessible en ligne sur `http://alpage.inria.fr/Lbx` (avec l'identifiant `guest` et mot de passe `guest`) et nous sommes ouverts au chargement de nouvelles ressources, éventuellement produites à partir de notre chaîne de traitement à partir d'un corpus spécialisé.

Références

AUBER D. (2003). Tulip : A huge graph visualisation framework. In P. MUTZEL & M. JÜNGER, Eds., *Graph Drawing Softwares*, Mathematics and Visualization, p. 105–126. Springer-Verlag.

BOURIGAULT D. (2002). UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Proc. of TALN'02*, p. 75–84, Nancy, France.

DEAN J. & GHEMAWAT S. (2004). MapReduce : Simplified data processing on large clusters. In *OSDI'04 : Sixth Symposium on Operating System Design and Implementation*, San Francisco, CA.

HARRIS Z. (1968). *Mathematical Structures of Languages*. New-York : John Wiley & Sons.

JACQUIN C., DESMONTILS E. & MONCEAUX L. (2007). French eurowordnet lexical database improvements. In *In Proc. of CICLing'07*, number 4394 in LNCS, Mexico City, Mexico.

KANERVA P., KRISTOFERSON J. & HOLST A. (2000). Random indexing of text samples for latent semantic analysis. In L. GLEITMAN & A. E. JOSH, Eds., *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, p. p. 1036, Mahwah, New Jersey : Erlbaum.

LIN D. & PANTEL P. (2002). Concept discovery from text. In *Proceedings of Conference on Computational Linguistics (COLING-02)*, p. pp. 577–583, Taipei, Taiwan.

PAZIENZA M. T., PENNACCHIOTTI M. & ZANZOTTO F. M. (2005). Terminology extraction : an analysis of linguistic and statistical approaches. In S. (ED.), Ed., *Knowledge Mining*, volume 185 of *Studies in Fuzziness and Soft Computing*. Springer Verlag.

SAGOT B. & FIŠER D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues. In *TALN 2008*, Avignon, France.

TURNEY P. D. (2002). Mining the web for synonyms : PMI-IR versus LSA on TOEFL. *CoRR*, **cs.LG/0212033**.

VAN DONGEN S. (2000). *Graph Clustering by Flow Simulation*. Phd thesis, University of Utrecht.

VILLEMONTE DE LA CLERGERIE E. (2010). Building factorized TAGs with meta-grammars. In *TAG+10 : The 10th International Conference on Tree Adjoining Grammars and Related Formalisms*, p. pp. 111–118, New Haven, CO.

VILNAT A., PAROUBEK P., VILLEMONTE DE LA CLERGERIE E., FRANCOPOULO G. & GUÉNOT M.-L. (2010). PASSAGE syntactic representation : a minimal common ground for evaluation. In *LREC*, La Vallete.

 © ATALA

Typologie et ressources pour le TALN des langues caucasiennes : le cas du tchétchène

Johannes Heinecke[1]
(1) Orange Labs, 2 avenue Pierre Marzin, 22307 Lannion Cedex
johannes.heinecke@orange.com

RÉSUMÉ

Le tchétchène est la deuxième langue caucasienne en terme de nombre de locuteurs. Avec une morphologie riche et une syntaxe complexe elle pose un certain challenge aux approches TALN. L'absence de ressources comme des corpus des textes électroniques (sauf 2000 articles courts de Wikipedia) interdit des approches statistiques. Après un survol sur les caractéristiques typologiques de cette langue ainsi que des problèmes quantitatifs et qualitatifs des corpus disponibles, nous proposons une analyse syntaxique du tchétchène basée sur une « Grammaire de dépendance à contraintes ».

ABSTRACT

Natural Language Processing for Caucasian Languages

The Chechen language is the second most spoken language in the linguistically extremely rich region of the Caucasus. Having a rich inflectional morphology as well as a complex syntax, Chechen poses a challenge to NLP approaches. The lack of computational linguistic resources and the virtual absence of electronical text corpora (except about 2000 short Wikipedia articles) make statistical approaches infructuous. After an overview of the typological characteristics of Chechen and addressing the quantitative and qualitative problems of the Wikipedia corpus, we propose a syntactic analysis of written Chechen based on « Constraint Dependency Grammars »

MOTS-CLÉS : Tchétchène, langues caucasiennes, encoding, ressources, TALN.

KEYWORDS: Chechen, Caucasian languages, encoding, NLP, resources.

1 Langues caucasiennes

Traditionnellement on entend par « langues caucasiennes » les langues qui appartiennent aux trois familles autochtones de la région du Caucase. Il s'agit de familles de langues caucasiennes du Nord-Ouest ou langues abkhazo-adygiennes, de langues caucasiennes du Nord-Est ou langues nakho-daghestaniennes et des langues caucasiennes du Sud ou langues kartvéliennes. Les premières deux familles sont parlées dans la partie nord du Caucase, appartenant à la Fédération Russe, les langues kartvéliennes dont le géorgien et l'abkhaze (une langue abkhazo-adygienne) au sud de la crête du Caucase, en Géorgie et en Abkhazie (cf. figure 1).

Les autres langues de la région caucasienne, qui ne font pas partie de ces trois familles (et qui sont très probablement génétiquement non liées entre elles), ne sont pas considérées comme langues caucasiennes, malgré leur présence dans la région : L'ossète (en Ossétie du Nord et Ossétie du

 © ATALA

FIGURE 1 – Carte des langues parlées dans la région caucasienne

Sud, langue indo-européenne, branche iranienne), le koumyk et le nogal (en Daghestan, langues turques), l'arménien (langue nationale de l'Arménie, langue indo-europénne) ou l'azeri (langue nationale de l'Azerbaïdjan, langue turque). Il n'est pas évident de dresser une liste exhaustive des langues caucasiennes, car certaines langues sont très proches (communication interlinguistique possible), mais les différences font que certaines écoles linguistiques parlent quand même de langues distinctes, d'autres préfèrent parler de variétés (ou dialectes) d'une même langue (Berg, 2005; Hewitt, 2005; Boeder, 2005).

Les langues caucasiennes ayant le plus grand nombre de locuteurs sont le géorgien avec 4 millions de locuteurs (recensement de la population 1989, Boeder (2005)), le tchétchène (langue de la branche Vaïnakh des langues nakho-daghestaniennes) avec environ un million ((Berg, 2005), recensement 1989), le kabarde (langue de la branche circassienne des langues abkhazo-adygiennes) avec 400 000 locuteurs (Hewitt (2005), récensement 1989) et l'avar (langue daghestanienne) (avec un nombre de locuteurs entre 500 et 600 000 si on inclut les dialectes proches (Berg, 2005)) [1]. Le dernier notamment a une fonction de *lingua franca*, avec le russe, dans la république Daghestan (Russie) où elle est aussi langue officielle (parmi d'autres).

Certaines de langues caucasiennes ne sont parlées que dans un village ou dans un arrondissement (notamment au Daghestan). D'autres comme le tchétchène ou le géorgien sont les langues maternelles de presque toute la population de la région (République Tchétchène dans la Fédération Russe ou la République Géorgienne). La distinction entre langue nationale et langue régionale n'as pas beaucoup de pertinence dans cette partie du monde, où presque tout le monde est parfaitement bilingue voir trilingue (y compris le russe). Aujourd'hui le géorgien est la langue nationale de la Géorgie qui a d'autres langues minoritaires sur son territoire. La République

1. Les recensements de 2002 (http://www.perepis2002.ru/index.html?id=87) et de 2010 (http://www.gks.ru/free_doc/new_site/perepis2010/croc/Documents/Vol4/pub-04-05.pdf) affichent des nombres de locuteurs bien plus élevés pour toutes les langues, par contre « connaissance de langue : oui » ne veut pas dire nécessairement « langue maternelle » ou « compétences parfaites ».

© ATALA

Tchétchène est une partie de la Fédération Russe, donc le russe y est aussi une langue officielle à coté du tchétchène. Est-ce que le tchétchène est une langue régionale (car il n'est pas parlé sur tout le territoire de la Russie) ou est-ce qu'il est une langue nationale (de la république Tchétchène) ? Le Daghestan, sur le territoire duquel la plupart des langues nakho-daghestaniennes sont parlées, possède une dizaine de langues comme officielle (dont le russe, l'avar et le tchétchène). Indépendamment si on considère ces langues comme nationales ou régionales, elles font toutes partie des langues « peu dotées » en ressources pour le TALN. C'est la raison pour laquelle nous examinons en détail la situation d'une des langues les plus parlées : le tchétchène.

Toutes les langues caucasiennes se distinguent par une phonologie et morphologie riches ainsi que par une syntaxe complexe. La construction ergative ou mixte dans le cas du géorgien rend par fois les approches linguistiques plus difficiles [2]. En général, les langues du Caucase possèdent beaucoup de traits typologiques, qui ne se trouvent pas dans les langues indo-européennes (Haspelmath, 2001).

1.1 Caractéristiques typologiques

Le tchétchène (*noxčijn mott*) constitue, avec l'ingouche, les langues Vaïnakh, un groupe des langues nakh dans la famille des langues nakho-daghestaniennes (Nichols, 1994). L'ingouche est très proche du tchétchène, les locuteurs des deux langues peuvent communiquer entre eux (La séparation du continuum tchétchéno-ingouche en deux langues (littéraires) différentes s'est établie au début du XXe siècle). La morphologie du tchétchène est complexe. Il dispose de 8 cas grammaticaux (tableau 1, cf. (Aliroev, 1999)) et 6 cas grammaticaux locatifs en plus (au lieu de prépositions locatives comme *à*, *de*, *en* ou encore *vers* en français). Dans certains cas l'ablaut peut modifier la voyelle de la racine du mot fléchi. Le tableau ne montre qu'une classe de déclinaison pour des raisons d'espace. Aux cas grammaricaux s'ajoutent les cas locatif, quis ont des noms très variés selon la littérature, au lieu d'utiliser un nom, nous rajoutons une traduction indicative.

	singulier	pluriel
ABS	*olxazar*	*olxazar-š*
GEN	*olxazar-an*	*olxazar-ijn*
ERG	*olxazar-ō*	*olxazar-š-a*
DAT	*olxazar-na*	*olxazar-š-na*
INST	*olxazar-ca*	*olxazar-š-ca*
COMP	*olxazar-al*	*olxazar-el*
MAT	*olxazar-ax*	*olxazar-ex*
LOC	*olxazar-ē*	*olxazar-š-ka*

cas locatifs	traduction indicative
bazar-ax, bazar-gaxula, bazar-exula	au delà (du marché)
bazar-axula	à travers
bazar-alla	vers, en direction vers
bazar-eḥara	àcoté de
bazar-garaḥara	vers

TABLE 1 – déclinaison nominale

Un trait intéressant du tchétchène (partagé par des autres langues nakho-daghestaniennes) sont ses 6 genres (ou classes, tableau 2). L'accord entre verbe et nom se fait uniquement par classe (et nombre). L'adjectif s'accorde avec le nom aussi en cas grammatical.

2. Pour un aperçu général des langues caucasiennes, voir (Boeder, 2005), (Berg, 2005) et (Hewitt, 2005)

3. (Deeters, 1963, p. 54) note qu'au pluriel *d* est utilisé pour la première et deuxième personne (pronoms) ; *b* pour la troisième personne (noms et pronoms).

© ATALA

classe	singulier	Pluriel	sémantique indicative (Deeters, 1963)
1	*v-*	*d-, b-*[3]	hommes
2	*j-*	*d-, b-*	femmes
3	*d-*	*d-*	enfants, objets naturels, classe non-marquée pour des référents inconnus (cf. allemand *es*), mots d'origine russe
4	*j-*	*j-*	plantes utiles, produits, métaux, mots d'origine arabe
5	*b-*	*b-*	expressions collectives
6	*b-*	*d-*	(aucune)

TABLE 2 – classes nominales

(1) *d-oq̇q̇* *d-i'* *ž'äla* *d-u*
`CL3.grand` `CL3.quatre` `chien.CL3` `CL3.SG.AUX`
« il y a quatre grands chiens »

v-oq̇q̇ *v-i'* *stag* *v-u*
`CL1.grand` `CL1.quatre` `homme.CL1` `CL1.SG.AUX`
« il y a quatre grands hommes »

(2) *gowr* *xi-x* *j-eli-ra*
`cheval.CL4.SG.ABS` `eau-MAT` `CL4.SG.aller-PERF`
« le cheval a traversé la rivière »

Le verbe dispose d'environ 10 temps grammaticaux ainsi que des morphèmes dérivatifs pour exprimer le causatif, le permissif ou la combinaison des deux (tableau 3[4], ici uniquement première classe de conjugaison ; voir (Jakovlev, 1960, p. 191), (Nichols, 1994, p. 46) pour plus d'informations sur la morphologie verbale, notamment les participes).

Certains verbes marquent le pluriel de l'actant en absolutif en plus de la classe (genre) par un changement de la racine (par ex ablaut : *lalla – liella*, verbe transitif : « envoyer quelqu'un une fois – plusieurs fois » ; *ḥaža – ḥōwsa*, verbe intransitif : « quelqu'un regarde – plusieurs (personnes) regardent »). Finalement l'opposition de l'aspect lexical[5] semelfactif (faire une fois) vs. itératif (faire plusieurs fois) peut se manifester dans des racines verbales différentes : *d*īġa – d*ōġa* « planter ». Vu le nombre des formes possibles à partir d'un lemme, des corpus plus importants sont nécessaires, afin de trouver certaines formes plusieurs fois.

Comme la plupart des langues caucasiennes, le tchétchène est aussi une langue ergative (l'agent d'un verbe intransitif porte l'absolutif mais l'agent d'un verbe transitif porte l'ergatif. Le verbe est en accord avec l'actant en absolutif (agent d'un verbe intransitif ou patient d'un verbe transitif, cf. (Bickel et Nichols, 2001; Guérin, 2010)) :

(3) *so* *ġalar-a* *v-ea*
`1SG.ABS.CL1` `ville-ILL` `CL1.venir-PRS`

4. « d* » indique que cette lettre change en fonction du genre de l'actant en absolutif.
5. En allemand « Aktionsart » ne pas à confondre avec l'opposition d'aspect accompli et inaccompli.

© ATALA

	verbe	causatif	permissif	permissif causatif	potential	inceptif
	« dire »	« faire dire »	« permettre dire »	« permettre faire dire »	« être capable de dire »	« commencer de dire »
infinitif	*āla*	*ālō*	*ālijta*	*ālad*ajta*	*ālad*ala*	*ālad*āla*
prés. simple	*ōlu*	*ālad*o*	*ōlüjti*	*ālad*ojta*	*ālalo*	*ālad*ōlu*
prét. proche	*ēli*	*ālij*	*ālijti*	*ālad*ajti*	*ālad*eli*	*ālad*ēli*
prét. temoigné	*ēlira*	*ālijra*	*ālijtira*	*ālad*ajtira*	*ālad*elira*	*ālad*ēlira*
prét. répété	*ōlura*	*ālad*ora*	*ōlüjtura*	*ālad*ojtura*	*ālalora*	
parfait	*älla*	*ālijna*	*ālijtina*	*ālad*ajtina*	*ālad*älla*	*ālad*älla*
plusqueparfait	*āllēra*	*ālijnēra*	*ālijtinēra*	*ālad*ajtinēra*	*ālad*elēra*	*ālad*āllēra*
futur possible	*ōlur*	*ālad*er*	*ōlüjtur*	*ālad*ojtur*	*ālalyr*	*ālad*olur*
futur réel	*ōlur d*u*					

TABLE 3 – temps grammaticaux du verbe tchétchène

« je (homme) viens à la ville »

so *ġalar-a* *j-ea*
`1SG.ABS.CL2` `ville-ILL` `CL2.venir-PRS`
je (femme) viens à la ville »

Musa *ċa-v-zġa* *v-eza*
`Musa.CL1.ABS` `maison-CL1-aller` `CL1-devrait`
« Musa devrait aller à la maison »

(4) *as* *kexat* *jaz-d-ina*
`1SG.ERG` `lettre.ABS.CL3` `écrire-CL3-PERF`
« j'ai (homme) écrit une lettre »

Musa-s *disertacii* *jaz-j-a* *j-eza*
`Musa.CL1-ERG` `thèse.CL4` `écrire-CL4-INF` `CL4-devrait`
« Musa devrait écrire (sa) thèse »

Ce sont notamment les marqueurs de classe infixés (exemple 4), l'ablaut des verbes ou noms (flexion) qui compliquent les approches s'appuyant sur le stemming. A notre connaissance, aucune système de stemming existe. Pour l'analyse syntaxique (section 3) nous utilisons un lexique full-forme.

2 Technologie d'information

2.1 Codage, localisation

Pour le tchétchène la situation est compliquée. La tradition écrite du tchétchène (la langue littéraire) est relativement récente. Les premiers textes rédigés en tchétchène (souvent de nature

 © ATALA

religieuse) utilisaient l'alphabet arabe, qui d'ailleurs pose des problèmes car le tchétchène a 8 phonèmes vocaliques et un grand nombre de phonèmes consonantiques, plus qu'il n'y a de graphèmes dans l'alphabet arabe. Après une courte période en alphabet latin au début du XXe siècle, l'alphabet cyrillique a été adopté en 1938 en URSS. Suite à la fin de l'URSS un nouveau alphabet latin à base de l'orthographe turque était proposé en 1992 (tableau 4 [6]). Cette alphabet latin est plutôt utilisé par la diaspora tchétchène, en Tchétchénie l'alphabet cyrillique est principalement utilisé.

I	*II*	*III*	*I*	*II*	*III*	*I*	*II*	*III*	*I*	*II*	*III*
а	a	a	к	k	k	с	s	s	чI	č̣	ç̣
аь	ä	ä	кх	q	q	т	t	t	ш	š	ş
в	v, w	v	къ	q̇	q̇	тI	ṭ	th	(щ)	ščč	
г	g	g	кI	ḳ	kh	у	u	u	(ъ)		’
гI	ǵ	ǵ	л	l	l	уь	ü	ü	ы	y	e
д	ḋ	ḋ	м	m	m	ф	f	f	(ь)		
е	e, je	e	н	n	n	х	x	x	э	e	e
(ё)	jo		о	o	o	хь	ḥ	x̣	ю	ju	yu
ж	ž	ƶ	оь	ö	ö	хI	h	h	юь	jü	yü
з	z	z	п	p	p	ц	c	c	я	ja	ya
и	i	i	пI	ṗ	ph	цI	ç	ċ	яь	jä	yä
й	j	y	р	r	r	ч	č	ç	I	ᶜ	j

TABLE 4 – L'alphabet tchétchène

Aujourd'hui les deux alphabets coexistent, notamment dans les ressources du Web on peut trouver parfois un mélange des deux (par exemple la page de Wikipedia tchétchène sur la langue tchétchène [7]).

Au problème que pose le mélange d'orthographe, si on travaille avec des corpus, s'ajoute celui concernant l'orthographe cyrillique : Le caractère « I » qui est utilisé dans des bigraphes [8] pour marquer la qualité éjective des plosives, pour exprimer le phonème /h/ <xI> ainsi en position isolée pour exprimer le fricatif pharyngal, existe sur trois positions différentes dans Unicode : U+0049 (latin capital letter I), U+0406 (cyrillic capital letter byelorussian-ukrainian I) et U+04C0 (cyrillic letter palochka). Pour le tchétchène en orthographe cyrillique on devrait utiliser uniquement le caractère U+04C0, mais la réalité des corpus (pages web, notamment Wikipedia) est différente et on trouve tous les trois caractères au lieu de la *palochka* et aussi le caractère U+006C (latin small letter l). Cela impose donc un prétraitement pour unifier le codage.

6. Dans la colonne I se trouve l'alphabet cyrillique de 1938, la colonne II montre la translittération scientifique de l'alphabet cyrillique (utilisé ici pour les exemples en langue tchétchène) enfin la colonne III contient l'alphabet latin de 1992. Les caractères entre parenthèses ne sont utilisés que dans des mots empruntés d'origine russe.

7. http://ce.wikipedia.org/wiki/%D0%9D%D0%BE%D1%85%D1%87%D0%B8%D0%B9_%D0%BC%D0%BE%D1%82%D1%82

8. Les bigraphes (voir tableau 4) sont considérés comme des lettres unitaires, par contre dans les supports électroniques il s'agit de deux caractères différents, car les ligatures correspondantes sont absentes dans Unicode ou des encodates à 8 bit pour le cyrillique.

© ATALA

2.2 Ressources électroniques

Il y a peu de ressources adéquates pour le tchétchène. Indépendamment des applications et des approches que l'on pourrait choisir (à base de règles ou statistiques), une ressource de base, qui est indispensable, sont des corpus. Grâce à Wikipedia [9] nous avons à notre disposition des corpus de textes (avec les petits problèmes cités ci-dessus) pour le tchétchène. Le tableau 5 montre les chiffres de Wikipedia tchétchène (et pour comparaison les chiffres des Wikipedia géorgien (langue avec un contexte comparable), basque (langue avec un nombre similaire de locuteurs) et français) [10].

	tchétchène	*géorgien*	*basque*	*français*
version du :	27.02.2013	26.02.2013	22.04.2013	16.02.2013
redirections :	657	26 255	49 035	1 281 842
pages spéciales :	2 929	47 781	125 766	516 082
pages normales :	1 939	72 388	150 509	1 352 801
page la plus longue (mots) :	19 983	11 737	16 068	94 422
page la plus longue (caractères) :	125 502	98 711	132 714	546 663
nombre moyen de mots par page :	261,21	133,75	190,74	384,54
nombre médian de de mots par page :	20	57	81	157
nombre moyen de caractères par page :	1 629,10	1064,67	1 458,07	2 387,24
nombre médian de caractères par page :	155	450	630	966

TABLE 5 – statistiques sur Wikipedia

La faible taille de la Wikipedia tchétchène (par rapport au géorgien ; toutes les deux sont petit par rapport aux Wikipedias français, anglais ou allemand) est déjà un indice fort sur la situation des corpus dans les langues avec un faible nombre de locuteurs. Il est difficile d'obtenir une masse critique de phrases en vue des apprentissages ou des évaluations. La situation empire à cause de la qualité de pages : beaucoup de pages sont très courtes ou simplement de listes (par ex. liste des « top level domains » de l'internet ou des pages d'évènements par jour de l'année) sans aucun vrai texte.

En revanche pour d'autres langues du Caucase, notamment le géorgien, on trouve des journaux en-ligne [11]. Pour le tchétchène par contre, a part du site en langue tchétchène de Radio Free

9. http://ce.wikipedia.org/

10. Ces chiffres incluent des citations en langue étrangère (souvent anglais ou russe) dans les pages. Les chiffres sur le nombre des mots ou caractères sont pris uniquement à partir des pages normales après avoir extrait le texte brut de la syntaxe wiki (avec l'aide de la librairie java http://sweble.org/). Les pages spéciales sont les pages dont le titre commence avec un namespace comme « MediaWiki » (définis au début du fichier dump XML des pages Wikipedia, http://dumps.wikimedia.org/cewiki/20130227/cewiki-20130227-pages-articles.xml.bz2).

11. Cf. l'archive des journaux *kartuli ḳresis elekṭronuli arkivi*, http://www.opentext.org.ge/ ou http://dir.geres.ge/en/

© ATALA

Europe, *Radio Maršo* [12] il n'y a pas de journal accessible sur le Web. La morphologie complexe du tchétchène a une autre conséquence : la proportion entre tokens et types est petite. Par exemple dans Wikipedia, nous comptons en total 404 647 mots (token, sans nombres, uniquement mots en alphabet cyrillique, majuscules et minuscules confondues) dont 41 511 mots différents (types) la relation token/type est donc 9,75. Chaque mot apparait donc en moyenne 9,75 fois dans tout le corpus, le mot le plus fréquent (*a* « et ») apparait 14 462 fois (3,6% des tokens), et 56% de tous les types n'apparaissent qu'une fois [13]. Ces chiffres montrent que pour beaucoup de approches statistiques (par exemple pour l'analyse syntaxique), actuellement les corpus disponibles sont insuffisants en taille.

Pour les ressources lexicales, la situation est similaire. Il existe des dictionnaires détaillés sur support papier [14], mais ils ne sont pas disponibles ou pas adaptés pour un traitement automatique. Il existe quelques dictionnaire en ligne [15], qui nécessiteront aussi une adaptation pour des approches TALN (par exemples formes fléchies).

Concernant la parole, il y a actuellement peu de stations radio ou TV tchétchènes sur internet (nous avons déjà mentionné le service en langue tchétchène par Radio Free Europe, Radio Marša, qui fourni en streaming audio en tchétchène), mais il est possible d'en trouver sur des sites comme youtube ou dailymotion ou bien des fichiers audio [16].

3 Tentative de génération d'arbres syntaxiques

Lors de notre travail sur le tchétchène qui a abouti à un dictionnaire électronique bilingue tchétchène - allemand (Heinecke, 1999a,b), nous avons développé un analyseur syntaxique afin de pouvoir plus facilement préparer des traductions interlinéaires des exemples (par ex. 1 à 4). Ce travail a été effectué afin de voir si l'outil de l'analyse est assez générique pour pouvoir traiter des langues typologiquement très différentes du « Standard Average European » (SAE) et afin de pouvoir produire plus facilement des traduction interlinéaires du tchétchène.

L'analyseur utilisé, qui produit des arbres de dépendances, s'appuie sur les « Constraint Dependency Grammars » (Maruyama *et al.*, 1990; Menzel et Schröder, 1998; Heinecke *et al.*, 1998; Heinecke et Schröder, 1998) et dispose d'une couche sémantique afin de prendre en compte des restrictions de sortes. Les grammaires consistent en contraintes syntaxiques avec des poids. En fonction des poids, certaines contraintes sont éliminées pour arriver à une solution. Par exemple l'analyse de 5 :

(5) *da-s* *kniga* *lō* *ḳant-ana.*
`père-ERG` `livre-ABS` `donner-PRES` `fils-DAT`
« le père donne le livre à son fils »

12. urlhttp ://www.radiomarsho.com/ ; ce site ne confond pas le caractère U+04C0 (cyrillic letter palochka) avec les autres caractère en apparence similaire, remarqué ci-dessus en section 2.1.
13. Pour comparaison Wikipedia français contient 327 173 000 token (sans nombres, chiffres, caractères non-latins, uniquement mots en alphabet latin, majuscules et minuscules confondues), dont 2 321 000 types (proportion 140,95). Le mot le plus fréquent « de » s'y trouve 22 292 000 fois, (6,8% des tokens)
14. (Maciev, 1961; Partchieva et Guérin, 1997; Aliroev, 2005; Nichols et Vagapov, 2004)
15. http://www.tsumada.ru/slovar, http://nohchiyn.net/sanscript/dosham2.php,
16. Cf. http://www.waynakh.com/eng/chechen-language/fairy-tales-in-chechen-language-mp3/

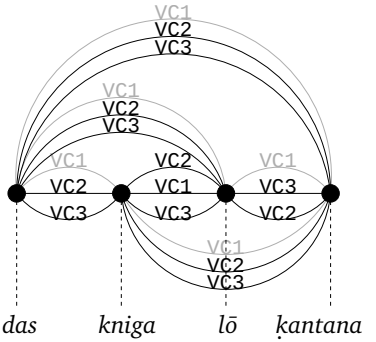
 © ATALA

Vu le fait que des catégories syntaxiques comme « sujet » ou « objet » ne sont pas applicable en tchétchène nous utilisons des catégories neutres VC*n* [17]. VC1 dénote le groupe nominale de nominatif (correspond au sujet ou objet dans la traduction française, en fonction de la valence du verbe), VC2 dénote le groupe nominale d'ergatif, et VC3 pour autre cas (à l'occurence le datif).

Initialement il y a une relation entre chaque nœud, les contraintes suppriment les relations incorrectes (marquées en gris clair dans les graphes).

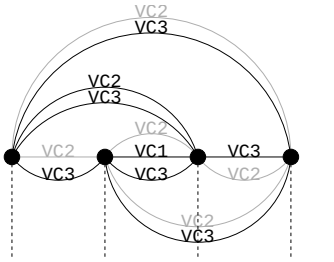
(6) {X:*SYN*} : *syn_vcompl1* : *syn_predicates* : 0.0 :
X.label = *VC1* % Si l'arête « X » a le libellé « VC1 »...
→ % implique
% au début de l'arête on trouve un verbe fini
X↑*cat* = *V-FIN*
% qui détermine le cas grammatical...
∧ X↓*case* = X↑*deps:dep1:case*
% ... la catégorie ...
∧ X↓*cat* = X↑*deps:dep1:cat*
% ... et la lasse du groupe nominal à la fin de l'arête
∧ X↓*class* = X↑*deps:dep1:class*

(7)



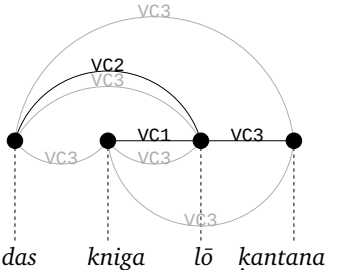
Des contraintes similaires suppriment toutes les arêtes jugées incorrectes :

(8) {X:*SYN*} : *syn_vcompl2* : *syn_predicates* : 0.0 :
X.label = *VC2* % Si l'arête « X » a le libellé « VC2 »...
→
X↑*cat* = *V-FIN*
∧ X↓*case* = X↑*deps:dep2:case*
∧ X↓*cat* = X↑*deps:dep2:cat*



(9)

(10) {X:*SYN*} : *syn_vcompl3* : *syn_predicates* : 0.0 :
X.label = *VC3*
→
X↑*cat* = *V-FIN*
∧ X↓*case* = X↑*deps:dep3:case*
∧ X↓*cat* = X↑*deps:dep3:cat*



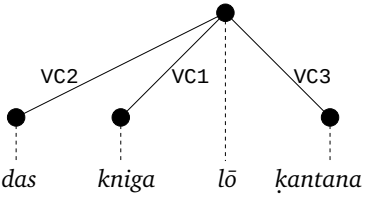
(11)

Et nous arrivons à la représentation syntaxique de la phrase (réarrangée sous forme d'arbre) :

17. Dans la littérature des langues ergatifs on utilise souvent les libellé S (sujet du verbe intransitif), A (sujet du verbe transitif), et O (objet du verbe transitif), mais cela nécessite de définir d'abord les critères pour « sujet » et « objet ». Or, en tchétchène on observe que le verbe s'accorde avec plusieurs groupes nominaux (cf. section 4 donc quel aurait le statut « sujet » et pourquoi pas l'autre ?

© ATALA

(12)



Afin de détecter des phrases syntaxiquement correctes mais peu plausibles sémantiquement, nous avons implémenté une deuxième série de contraintes sur un niveau sémantique (sortes). Les deux exemples suivants sont tous les deux syntaxiquement correctes, mais l'exemple 14 est improbable :

(13) *Ḳanta* `garçon-ERG` *döšu* `CL3-lire-PRES` *kēxat.* `lettre.CL3` « le garçon lit la lettre »

(14) **traktorō* `tracteur-ERG` *döšu* `CL3-lire-PRES` *kēxat.* `lettre.CL3` « *le tracteur lit la lettre »

Les contraintes de sélection sont codées dans le lexique :

(15) a) *döšu* :

$$\begin{bmatrix} \text{cat :} & \textit{finite verb} \\ \text{tense :} & \textit{present} \\ \text{class :} & \textit{3 (d)} \\ \text{deps :} & \begin{bmatrix} \text{dep1 :} & \begin{bmatrix} \text{role :} & \textit{THEME} \\ \text{sort :} & \textbf{\textit{any}} \\ \text{case :} & \textit{abs} \end{bmatrix} \\ \text{dep2 :} & \begin{bmatrix} \text{role :} & \textit{AG} \\ \text{sort :} & \textbf{\textit{human}} \\ \text{case :} & \textit{erg} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

b) *ḳanta* :

$$\begin{bmatrix} \text{cat :} & \textit{noun} \\ \text{sort :} & \textbf{\textit{human}} \\ \text{class :} & \textit{1 (v)} \\ \text{case :} & \textit{erg} \\ \text{number :} & \textit{sg} \end{bmatrix}$$

c) *traktorō* :

$$\begin{bmatrix} \text{cat :} & \textit{noun} \\ \text{sort :} & \textbf{\textit{object}} \\ \text{class :} & \textit{3 (d)} \\ \text{case :} & \textit{erg} \\ \text{number :} & \textit{sg} \end{bmatrix}$$

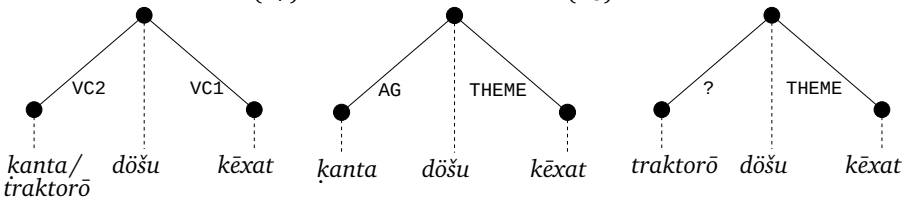
d) *kēxat* :

$$\begin{bmatrix} \text{cat :} & \textit{noun} \\ \text{sort :} & \textbf{\textit{object}} \\ \text{class :} & \textit{3 (d)} \\ \text{case :} & \textit{abs} \\ \text{number :} & \textit{sg} \end{bmatrix}$$

L'analyse syntaxique donne évidemment une structure identique (16) pour les deux phrases, par contre l'analyse sémantique de 14 n'aboutit pas (18) :

© ATALA

(16) (17) (18)



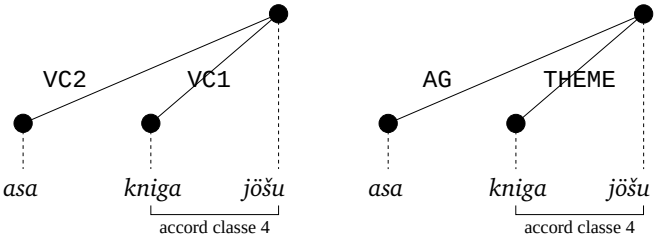
Dans notre cas, les contraintes sémantiques ont un petit poids, pour ne pas bloquer l'analyse syntaxique, notamment en absence d'information dans le lexique. (Heinecke, 1999b).

3.1 Verbes composés

En tchétchène les temps grammaticaux composés (comme le passé composé français) possèdent un trait intéressant concernant l'accord. Au lieu de s'accorder uniquement avec l'absolutif (19), l'agent et comme le patient marqué par l'absolutif, le verbe continue de s'accorder avec le patient, mais l'agent est en accord avec l'auxiliaire (21) :

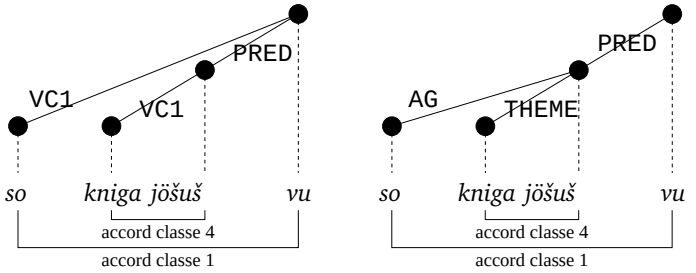
(19) *asa* `je-CL1.ERG` *kniga* `CL4.livre-ABS` *jöšu.* `CL4-lire-PRES`
« je (homme) lis le livre »

(20)



(21) *So* `je-ABS.CL1` *kniga* `CL4.livre-ABS` *jöšuš* `CL4.lire-GERUND` *vu.* `CL1-AUX-PRES`
« je (homme) suis en traint de lire un livre »

(22)

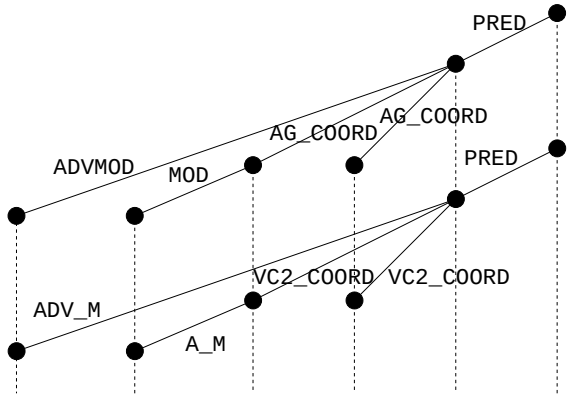


La coordination est un problème non-trivial pour les analyses de dépendance (par ex. « je mange des pommes et toi des poires »). Pour notre grammaire, nous ne traitons pas, pour

© ATALA

l'instant, la coordination simple de deux groupes nominaux. Un nœud spécial pourrait indiquer la coordination des nominaux ci-dessous :

(23) *cq̇a* *q̇ilbasēdera-ču* *mox-ō-j* *mälx-ō-j* *q̇ewsina* *xill-ēra...*
`une fois` `nord-OBL` `vent-ERG-et` `soleil-ERG-et` `se battre` `AUX-PRET`
« le vent du nord et le soleil se battaient » (Hewitt, 1978)



(24) *cq̇a* *q̇ilbasēderaču* *moxōj* *mälxōj* *q̇ewsina* *xillēra*

4 Problèmes et perspectives

Il y a des indices qui laissent supposer que le tchétchène dispose de l'accord multiple (plusieurs actants sont en accord avec le verbe, à part les verbes composés comme vu en section 3.1) comme le montre (Hewitt, 1978)

(25) *cundela* *q̇ilbasēderaču* *mexan* *juxabāla* *dijzira*
`donc` `nord-OBL` `CL6.vent-GEN` `CL4.CL6.renoncer-INF` `nécessaire-PRET`
šēn *ᶜālašōnan* *ṭēra.*
`POSS-3SG` `CL4.but-PL-GEN` `de.`
« donc le vent du nord devait renoncer à ces buts »

D'après Hewitt, *juxabāla* (*j-uxa-b-āla*) s'accorde avec les dexu noms, *mekhan* et *ᶜālašōnan*. Un fait suprenant est que les noms sont au génitif. L'exemple prototypique de (Maciev, 1961) n'affiche pas ce comportement :

(26) *eskar* *d-uxa-d-älla.* « l'armée s'est retirée »
`CL3-armee-ABS` `CL3-CL3-retirer-PRET`

Plusieurs locuteurs natifs m'ont confirmé qu'ils considèrent le préfixe verbal *juxa* comme non-modifiable. Donc ils acceptent des phrases comme 27.

(27) *so* *juxa-v-älla.* « je me suis retiré »
`CL1.je-ABS` `CL1.retire-PRET`

 © ATALA

Les mêmes personnes acceptent quand même l'exemple 26 avec le préfixe *duxa-*. Il est donc plus probable que les deux, *juxa* et *duxa* soient deux formes d'un même préfixe (locatif) sans aucun accord de classe. [18]

La version utilisée du Constraint Dependency Parser n'est pas optimale pour se mettre à l'échelle. C'est-à-dire qu'avec le lexique complet et toutes les contraintes nécessaires (environ 140) pour analyser le tchétchène écrit, les temps du traitement deviennent trop couteux. D'autres essais avec la boîte à outils freeling [19] ont l'air plus prometteurs, mais il est trop tôt pour en tirer des conclusions.

Le lexique utilisé contient environ 2000 noms, 400 adjectifs et 830 verbes (et les mots grammaticaux). Le but principal de ces travaux n'est pas non plus d'avoir un analyseur rapide, mais un mécanisme de support pour la création des traductions interlinéaires et de voir comment on peut traiter une langue sous dotée en ressources de TALN.

Abréviations

$\rightarrow$	implication	GERUND	gérondif
$\wedge$	« et » logique	INST	instrumentatif
$\vee$	« ou » logique	LOC	locatif
$=$	égal	MAT	matériau
$\neq$	différent	OBL	cas oblique
ABS	absolutif	PL	pluriel
AG	agent	PRES	temps présent
AUX	auxiliaire	PRET	temps passé
CL*x*	classe *x* (x = *1, 2, 3, 4, 5, 6*)	SG	singulier
COMP	comparatif	THEME	thème
COORD	coordination	VC*n*	complément verbal *n*
DAT	datif	V-FIN	verbe fini
DEMPR	démonstratif	V-INF	infinitif
ERG	ergatif	X↑*trait*	*trait* à la fin de l'arête « X »
ESS	(in)essif	X↓*trait*	*trait* au début de l'arête « X »
EXP	expérient		
GEN	génitif		

Références

ALIROEV, I. J. (1999). *Čečenskij Jazyk*. Yazyki narodov Rossii. Academia, Moskva.

ALIROEV, I. J. (2005). *Čečensko-russkij slovar'*. Izdatel'stvo Academia, Moskva.

BERG, H. v. d. (2005). The East Caucasian language family. *Lingua*, 115:147–190.

BICKEL, B. et NICHOLS, J. (2001). Syntactic ergativity in light verb complements. *In Syntactic ergativity in light verb complements*.

18. J. Nichols, communication personelle.
19. http://nlp.lsi.upc.edu/freeling/

© ATALA

BOEDER, W. (2005). The South Caucasian languages. *Lingua*, 115:5–89.

DEETERS, G. (1963). *Armenische und Kaukasische Sprachen*. Handbuch der Orientalistik : Abt. 1, Der Nahe und der Mittlere Osten, Bd. 7. Brill, Leiden.

GUÉRIN, F. (2010). Dynamique syntaxique en tchétchène. *Bulletin de la Société de Linguistique de Paris*, 105:1:247–263.

HASPELMATH, M. (2001). The European linguistic area : Standard Average European. *In* HASPELMATH, M., KÖNIG, E., OESTERREICHER, W. et RAIBLE, W., éditeurs : *Language Typology and Language Universals. Vol 1-2*, pages 1492–1510. de Gruyter, Berlin.

HEINECKE, J. (1999a). Computer Supported Morphological Database for Chechen. *In* BERG, H. v. d., éditeur : *Studies in Caucasian linguistics*, pages 254–266. CNWS, Leiden.

HEINECKE, J. (1999b). *Parsing Rules for Chechen. Paper given at Chicago Conferene on Caucausia*. Chicago.

HEINECKE, J., KUNZE, J., SCHRÖDER, I. et MENZEL, W. (1998). Eliminative Parsing with Graded Constraints. *In 17th CoLing*, pages 526–530. Université de Montréal, Montréal.

HEINECKE, J. et SCHRÖDER, I. (1998). Multilevel Representation of the Robust Analysis of Language. *In* SCHRÖDER, B., LENDERS, W. et HESS, W., éditeurs : *Konvens '98*, pages 13–17. Lang, Frankfurt.

HEWITT, B. G. (1978). The North Wind and the Sun in Chechen. Unpublished manuscript.

HEWITT, B. G. (2005). North West Caucasian. *Lingua*, 115:91–145.

JAKOVLEV, N. F. (1960). *Morfologija čečenskogo jazyka*. Čečeno-ingušskoe knižnoe izdatel'stvo, Groznyj.

MACIEV, A. G. (1961). *Čečensko-Russkij Slovar'*. Gosudarstvennoe Izdatel'stvo Inostrannyx i Nacional'nyx Slovarej, Moskva.

MARUYAMA, H., WATANABE, H. et OGINO, S. (1990). An Interactive Japanese Parser for Machine Translation. *In 13th CoLing*, pages 257–262. Helsinki.

MENZEL, W. et SCHRÖDER, I. (1998). Decision Procedures for Dependency Parsing Using Graded Constraints. *In* KAHANE, S. et POLGUÈRE, A., éditeurs : *Workshop of Processing of Dependency-based Grammars, 17th CoLing*, pages 78–87. Université de Montréal, Montréal.

NICHOLS, J. (1994). Chechen. *In* SMEETS, R., éditeur : *The indigenious languages of the Caucasus, Vol 4. The North East Caucasian languages II*, pages 3–77. Caravan Books, Delmar, NY.

NICHOLS, J. et VAGAPOV, A. (2004). *Chechen-English and English-Chechen Dictionary*. Routledge, London.

PARTCHIEVA, P. et GUÉRIN, F. (1997). *Parlons tchétchène-ingouche*. L'Harmattan, Paris.

© ATALA

Étiquetage morphosyntaxique de langues non dotées à partir de ressources pour une langue étymologiquement proche

Yves Scherrer Benoît Sagot
Alpage, INRIA & Université Paris-Diderot, 75013 Paris
{prenom.nom}@inria.fr

RÉSUMÉ

Nous présentons une approche générique du transfert d'annotations morphosyntaxiques d'une langue dotée vers une langue non dotée étymologiquement proche. Nous ne présupposons aucun corpus parallèle et aucune connaissance préalable de la langue non dotée (ni lexique, ni corpus annoté). Notre approche repose uniquement sur des paires de cognats obtenues par apprentissage non-supervisé selon le paradigme de la traduction automatique statistique à base de caractères, et sur un dictionnaire morphosyntaxique de la langue dotée. Pour les mots fréquents et courts, nous préférons assigner les étiquettes directement aux mots de la langue non dotée en fonction de mesures de similarité inter-langues du contexte morphosyntaxique immédiat. Partant de l'allemand comme langue dotée, nous évaluons notre approche sur le néerlandais, qui est en réalité dotée, et le palatin. Nous obtenons une précision d'étiquetage de 67,2% pour le néerlandais et de 60,7% pour le palatin.

ABSTRACT

Morphosyntactic tagging of non-resourced languages based on resources for another etymologically related language

We introduce a generic approach for transferring part-of-speech annotations from a resourced language to a non-resourced but etymologically close language. We do not rely on the existence of any parallel corpora or any linguistic knowledge for the non-resourced language (no lexicons, no annotated corpora). Our approach only makes use of cognate pairs that are automatically induced in an unsupervised way, based on character-based statistical machine translation and on a morphosyntactic lexicon for the resourced language. Frequent and short words are treated differently, as we tag them directly based on a cross-language similarity assessment of immediate morphosyntactic contexts. Using German as a resourced language, we evaluate our approach on Dutch — in fact a resourced language — and on Palatine German. We reach tagging accuracies of 67.2% on Dutch and 60.7% on Palatine German.

MOTS-CLÉS : étiquetage morphosyntaxique, langues proches, langues non-dotées.

KEYWORDS: part-of-speech tagging, etymologically close languages, non-resourced languages.

1 Introduction

Le traitement automatique des langues régionales pose un certain nombre de difficultés. En général, seulement une quantité limitée de textes écrits est disponible, et ces textes ne sont pas annotés. Mais souvent, il existe une langue mieux dotée en ressources qui est étymologiquement

 © ATALA

proche ; en effet, de nombreuses langues régionales d'Europe peuvent être vues — considérations politiques mises à part — comme des dialectes des « grandes » langues standardisées. Notre travail se situe dans ce type de configuration linguistique, où on observe à la fois une asymétrie quantitative des données (peu de données dans la langue régionale, beaucoup de données dans la langue standardisée) et une asymétrie qualitative (données annotées dans la langue standardisée, mais pas dans la langue régionale).

Nous présentons ici une approche générique du transfert d'annotations morphosyntaxiques (parties du discours) d'une langue dotée (ci-après, LD) vers une langue non dotée (LND), où les deux langues sont étymologiquement proches. Ce dernier point nous permet de formuler deux hypothèses quant à la similarité des deux langues. Premièrement, au niveau lexical, les deux langues partagent beaucoup de cognats, c'est-à-dire des paires de mots qui sont formellement (graphémiquement) et sémantiquement similaires. Deuxièmement, au niveau structurel, nous admettons que les systèmes de morphologie flexionnelle des deux langues sont similaires et que l'ensemble des catégories morphosyntaxiques est le même. Ainsi, nous supposons qu'à chaque fois que nous rencontrons une paire de cognats, la ou les catégories associées à l'un des mots peuvent être transférées à l'autre. Pour les mots fréquents et courts, que nous qualifierons abusivement de « mots grammaticaux », cette approche ne donne pas de résultats fiables ; dans ce cas, nous préférons assigner les étiquettes directement aux mots de la langue non dotée selon la similarité du contexte morphosyntaxique immédiat. Le texte ainsi annoté peut ensuite être utilisé comme corpus d'entraînement pour un étiqueteur.

Notre tâche est donc très différente de ce que l'on appelle en général l'*étiquetage morphosyntaxique non supervisé*. En effet, on trouve dans la littérature, depuis Merialdo (1994), de nombreux articles qui définissent sous ce nom la tâche consistant à effectuer un étiquetage morphosyntaxique sans corpus d'entraînement mais grâce à un lexique morphosyntaxique préexistant. Autrement dit, on projette un lexique sur un corpus brut, et la tâche revient alors à désambiguïser l'annotation ainsi obtenue. Dans la plupart des cas, des techniques d'apprentissage automatique sont mises en place afin d'induire un modèle probabiliste (cf. cependant Brill, 1995). Le modèle le plus populaire pour cette tâche est celui des Modèles de Markov Cachés (HMM) (Merialdo, 1994; Goldwater et Griffiths, 2007; Goldberg *et al.*, 2008; Ravi et Knight, 2009), mais d'autres modèles ont également été proposés, et notamment des modèles discriminants (Smith et Eisner, 2005). Différentes techniques d'apprentissage ont été utilisées, et des connaissances linguistiques (en plus du lexique) sont parfois intégrées, notamment pour initialiser les paramètres du modèle HMM.

À l'inverse de ces travaux, nous ne présupposons ici la disponibilité d'aucune ressource pour la LND, mis à part des données textuelles brutes. Nous évaluons notre approche sur des données de différentes langues germaniques. Nous avons choisi l'allemand comme langue dotée, et le néerlandais ainsi que le dialecte palatin [1] comme langues non dotées. La configuration allemand–néerlandais nous a servi pour le paramétrage du modèle, puisque des ressources annotées existent pour le néerlandais. La configuration allemand–palatin est plus proche du champ d'application type de notre méthode, mais nous disposons seulement d'un petit corpus annoté pour l'évaluation.

1. Le palatin (en allemand, *Pfälzisch*), ou francique rhénan du Palatinat, est un dialecte moyen-allemand parlé entre Heidelberg, Mannheim, Kaiserslautern et la frontière française. Il se distingue de l'allemand standard par des différences phonétiques régulières notamment au niveau consonantique, par la simplification de certains paradigmes flexionnels, et par certains phénomènes caractéristiques du langage parlé comme l'élision et la cliticisation. En revanche, les différences lexicales sont limitées.

 © ATALA

2 État de l'art

La complexité de notre tâche, due essentiellement à une disponibilité de ressources extrêmement contrainte, nécessite le recours à une combinaison originale de travaux de plusieurs domaines de recherche. Nous recensons ici les travaux desquels nous nous sommes inspirés, ainsi que de certains travaux connexes mais dont nous ne pouvions faire un usage pratique.

2.1 Détection de cognats

Hauer et Kondrak (2011) définissent les cognats comme des mots de langues différentes, mais avec une origine linguistique commune. Deux mots entrant dans une paire de cognats doivent satisfaire les trois propriétés suivantes : ils doivent être phonétiquement ou graphémiquement similaires, ils doivent être sémantiquement similaires, et les correspondances phonétiques ou graphémiques doivent être régulières. Cependant, la plupart des travaux se focalisent sur les deux critères de similarité phonétique ou graphémique.

En général, plus les langues en question sont étymologiquement proches, plus le pourcentage de cognats dans leur vocabulaire est élevé. La détection de cognats est donc une tâche peu coûteuse mais très utile pour notre travail. Un de nos objectifs est de détecter des paires de cognats pour en faire un lexique de traduction, nécessairement bruité. Cette tâche d'induction lexicale a été appliquée par Mann et Yarowsky (2001) : ils évaluent différentes mesures de distance phonétique ou graphémique. Ils distinguent notamment les mesures statiques (indépendantes de la paire de langues) des mesures adaptatives (adaptées à la paire de langues par apprentissage automatique). Sans surprise, les auteurs constatent de meilleures performances avec les mesures adaptatives. Hélas, ces mesures nécessitent un corpus d'apprentissage bilingue dont nous ne disposons pas.

Kondrak et Dorr (2004) présentent une multitude de mesures de distance ou de similarité qui ont l'avantage d'être complètement indépendantes de la langue et donc de ne pas avoir besoin d'un corpus d'apprentissage. Leur tâche est cependant un peu différente : ils veulent mesurer la distance entre des noms de médicaments pour prédire si un nouveau nom de médicament peut prêter à confusion avec des médicaments existants. Parmi les mesures qui opèrent sur des chaînes orthographiques (ils proposent également des mesures pour des chaînes phonétiques, mais qui ne nous intéressent pas ici faute de ressources transcrites phonétiquement), l'algorithme BI-SIM (voir section 3.1.1) donne les meilleurs résultats. Inkpen *et al.* (2005) appliquent ces mesures à la tâche d'identification de cognats dans des langues « proches » (anglais–français), et constatent notamment que des classifieurs supervisés ne font pas mieux que des mesures indépendantes de la langue avec un seuil judicieusement choisi.

2.2 Induction de lexiques à partir de corpus comparables

Les langues régionales d'Europe sont souvent dans une situation de diglossie, et leurs locuteurs maîtrisent presque toujours la langue à grande diffusion environnante. La demande de données traduites, et donc l'existence de corpus parallèles, s'en trouvent fortement réduites. En l'absence de corpus parallèles, on peut naturellement se tourner vers la recherche sur l'induction de lexiques à partir de corpus comparables. Cette ligne de recherche a été inaugurée indépendamment par Fung (1998) et Rapp (1999), dont l'idée consiste à examiner les contextes d'un petit

 © ATALA

ensemble de mots-amorces. L'hypothèse est alors qu'un mot français apparaissant souvent dans le contexte du mot *école* sera traduit par un mot anglais qui apparaît souvent à côté du mot *school*. L'inconvénient de cette procédure est le besoin d'une liste de mots-amorces traduits (dans notre exemple, la paire *école–school*), et le besoin de grands corpus pour les deux langues afin de pouvoir construire des vecteurs de similarité suffisamment larges.

Le premier point a été adressé par Fišer et Ljubešić (2011) pour des langues proches : ils utilisent des mots identiques et similaires (selon la mesure BI-SIM) pour construire automatiquement la liste des mots-amorces. De plus, ils disposent de corpus lemmatisés et étiquetés pour les deux langues, ce qui réduit la complexité de la tâche. Malheureusement, cette approche n'est pas adaptée à notre tâche, en raison de la petite taille du corpus et de l'absence d'outils d'analyse pour la LND — notre objectif est précisément de créer ce type d'outils.

Finalement, Koehn et Knight (2002) combinent différentes heuristiques, dont la similarité contextuelle et la similarité formelle des mots, pour induire un lexique de traduction à partir de corpus comparables.

2.3 Traduction automatique statistique à base de caractères

Le principe de la traduction automatique statistique (TAS) consiste à apprendre des alignements entre paires de mots apparaissant ensemble dans un corpus parallèle. Dans la TAS à base de segments, certaines paires de mots sont regroupées en paires de segments, ce qui a permis d'améliorer la performance (Koehn *et al.*, 2003). Récemment, une variation de ce modèle a vu le jour sous la forme de la TAS à base de caractères (Vilar *et al.*, 2007; Tiedemann, 2009). Dans ce paradigme, au lieu d'aligner des mots (et des segments de mots) dans un corpus consistant en phrases, on aligne des lettres (et des segments de lettres) dans un corpus consistant en mots. Evidemment, cette approche fait sens uniquement pour des cognats, où l'alignement des caractères est défini. Elle a donc été utilisée pour la traduction entre langues proches (Vilar *et al.*, 2007; Tiedemann, 2009), ainsi que pour la translitération (Tiedemann et Nabende, 2009). Dans ce dernier papier, les auteurs montrent notamment que la TAS à base de caractères donne de meilleurs résultats que des transducteurs probabilistes tels que ceux proposés par Mann et Yarowsky (2001) pour l'induction lexicale.

Ce modèle nécessite donc comme corpus d'apprentissage un ensemble de paires de mots. [2] Si dans les travaux existants ces données sont extraites de corpus parallèles, nous avons choisi de construire un corpus d'apprentissage de manière automatique en utilisant des mesures de similarité statiques telles que BI-SIM. Nous utiliserons la pipeline standard de la TAS : GIZA++ (Och et Ney, 2003) pour l'alignement des lettres, et de Moses (Koehn *et al.*, 2007) pour l'extraction de segments et le décodage (c'est-à-dire la traduction d'un mot source vers un mot cible).

2.4 Transfert d'annotations morphosyntaxiques

Notre objectif étant de transférer des étiquettes morphosyntaxiques d'une langue dotée vers une langue non dotée, une idée simple serait d'utiliser un corpus parallèle aligné par mots, et

2. On peut également modéliser les frontières des mots, auquel cas on utiliserait des segments de mots ou des phrases courtes comme données d'apprentissage. Cependant, nous nous limitons à des paires de mots simples dans nos expériences.

© ATALA

de copier l'étiquette des mots de la LD vers les mots alignés de la LND. Cette méthode a été proposée par Yarowsky *et al.* (2001) et appliquée à l'étiquetage morphosyntaxique, au chunking de syntagmes nominaux, à l'annotation d'entités nommées et même à l'induction de paradigmes morphologiques. Pour toutes ces tâches, les auteurs observent que les annotations transférées sont très bruitées, notamment à cause des erreurs d'alignement. Ils arrivent à éliminer une partie du bruit en entraînant un étiqueteur sur le corpus transféré et en ré-étiquetant le corpus de la LND. Nous nous inspirons de ce travail, mais l'absence de corpus parallèles nous incite à utiliser des alignements obtenus par similarité graphémique.

Une autre approche de la même problématique a été proposée par Feldman *et al.* (2006). Ils entraînent un étiqueteur sur la LD et l'appliquent, moyennant quelques modifications, à la LND. Évidemment, un tel étiqueteur va rencontrer de nombreux mots de la LND qu'il ne saura pas étiqueter. Feldman *et al.* (2006) poursuivent deux stratégies pour pallier à ce problème. Premièrement, ils utilisent un analyseur morphologique basique pour la LND, écrit à la main, pour prédire les étiquettes potentielles parmi lesquelles l'étiqueteur devra en choisir une. Deuxièmement, ils induisent une liste de cognats à l'aide d'une distance d'édition manuellement adaptée à la paire de langues, qui vont permettre à l'étiqueteur de choisir la catégorie du mot de la LD correspondant. Cette approche est prometteuse, mais n'est pas adaptée à notre contexte pour deux raisons. D'abord, il est difficile de créer un analyseur morphologique, même basique, pour certaines langues ou dialectes non standardisés. Ensuite, selon l'étiqueteur choisi, il peut être difficile de modifier les paramètres internes d'un modèle une fois la phase d'entraînement terminée.

3 Expériences

Comme évoqué dans l'introduction, nous proposons une approche inédite, à notre connaissance, pour transférer des annotations morphosyntaxiques d'une langue dotée (LD) vers une langue non dotée (LND) sans requérir la disponibilité de corpus parallèles. Concrètement, nous partons donc de ressources que l'on peut catégoriser en trois types :

1. Un texte brut en LND qui sera annoté à la fin du processus. Comme évoqué en introduction, nous distinguons dans ce corpus les « mots grammaticaux », définis par convention comme les 200 mots les plus fréquents ayant moins de 6 caractères, des autres mots, dits « mots pleins ». Outre des informations fréquencielles par mot-type, nous extrayons de ce corpus les contextes morphosyntaxiques des mots grammaticaux, en vue d'en prédire la catégorie.
2. Un texte brut de gros volume en LD, dont nous extrayons également la liste des mots-types avec leurs fréquences d'occurrence.
3. Un corpus annoté en parties du discours en LD, utilisé pour extraire un lexique morphologique de base ainsi que les contextes morphosyntaxiques des mots grammaticaux.

Nous appliquons notre méthode aux paires néerlandais–allemand et palatin–allemand. Pour les textes bruts, nous avons choisi des textes de Wikipedia : la Wikipedia entière pour le palatin et l'allemand, et un sous-ensemble de 500 000 mots pour le néerlandais, afin de créer des conditions comparables à l'expérience palatin–allemand. En ce qui concerne le corpus annoté allemand, nous utilisons TIGER (Brants *et al.*, 2002). Ce dernier est annoté avec le jeu d'étiquettes STTS (Stuttgart-Tübingen Tagset) qui comporte 54 étiquettes (Thielen *et al.*, 1999). La table 1 résume la taille de ces données. Pour l'évaluation de notre méthode (cf. section 4), nous utiliserons

 © ATALA

Type	Source	Langue	Mots	Mots-types
1	Wikipedia	Palatin	272 111	51 002
1	Wikipedia	Néerlandais	500 151	57 917
2	Wikipedia	Allemand	522 934 377	9 347 129
3	TIGER	Allemand	885 067	85 163

TABLE 1 – Données utilisées

comme référence la Wikipedia néerlandaise annotée à l'aide d'un étiqueteur classique, et un fragment de 100 phrases de la Wikipedia palatine annotée manuellement.

Le modèle que nous proposons sépare le traitement des « mots pleins » de celui des « mots grammaticaux ». Pour les mots pleins, nous essayons de trouver des paires de cognats en LND et LD, et de transférer les étiquettes morphosyntaxiques connues des mots LD vers les mots LND correspondant (sections 3.1 et 3.2). Pour les mots grammaticaux, souvent courts et comportant des correspondances graphémiques irrégulières, la détection de cognats est plus difficile. Au lieu d'induire des paires de cognats, nous essayons donc de prédire leurs étiquettes à l'aide de leurs contextes morphosyntaxiques dans nos ressources textuelles (section 3.3).

3.1 Création du lexique ouvert par alignement de lettres

Pour formaliser la problématique, nous définissons une **paire de cognats catégorisée (PCC)** $(w_{\text{LND}}, w_{\text{LD}}, c)$ comme une entrée lexicale bilingue telle que le mot w_{LND} (de la langue non dotée) et le mot w_{LD} (de la langue dotée) sont des cognats, et que les deux mots ont la catégorie morphosyntaxique c.

Nous créons le lexique de PCC uniquement pour les mots « pleins », et cela en plusieurs phases. D'abord, on récolte, pour chaque w_{LND}, l'ensemble des w_{LD} qui sont le plus similaires, selon une mesure de similarité graphémique indépendante de la paire de langues (section 3.1.1). Ensuite, ces paires de mots sont utilisés pour entraîner un modèle de traduction automatique basé sur les caractères (section 3.1.2). Il s'ensuit une étape de filtrage des résultats ainsi obtenus (section 3.1.3). À ce moment, nous disposons de paires de cognats non catégorisées $(w_{\text{LND}}, w_{\text{LD}})$. Nous rajoutons les informations de catégorie à partir d'un ensemble de paires (w_{LD}, c) extraits du corpus annoté de la langue dotée (section 3.1.5).

Pour des raisons pratiques, nous inférons les paires de cognats dans le sens $w_{\text{LND}} \rightarrow w_{\text{LD}}$. Cela nous permet notamment d'obtenir le même w_{LD} pour plusieurs w_{LND} différents et ainsi de tenir compte de variations orthographiques dans la LND. De telles variations sont moins attendues dans la LD, qui en principe est standardisée. De plus, il est préférable de créer le modèle de langage pour la LD, suivant la distribution quantitative de nos données ; selon l'architecture de la TAS, le modèle de langage doit être situé du côté de la langue cible.

3.1.1 Extraction de paires de cognats potentiels

Nous commençons par créer des listes de mots des deux corpus extraits de Wikipedia. En faisant cela, nous ne tenons pas compte des mots courts ($\leq$ 4 caractères), ni des hapax. De plus,

 © ATALA

pour l'allemand, nous nous limitons aux mots apparaissant plus de 1000 fois dans le corpus Wikipedia. [3] Ces filtrages nous donnent 20 225 mots néerlandais, 13 466 mots palatins et 26 951 mots allemands. [4]

La similarité formelle entre deux mots est calculée à l'aide de la mesure BI-SIM (Kondrak et Dorr, 2004). BI-SIM est une mesure de similarité graphémique qui utilise des bigrammes de caractères comme unités de base. Elle ne permet pas d'alignements croisés, et elle est normalisée selon la longueur de la chaîne la plus longue. Ainsi, elle permet de capturer un certain degré de sensitivité au contexte, évite de postuler des alignements non-intuitifs et favorise des associations entre mots de longueur similaire. L'avantage de la mesure BI-SIM est qu'elle est complètement générique et qu'elle ne présuppose donc aucune connaissance des relations étymologiques entre les deux langues. En revanche, elle n'est pas très précise et génère beaucoup d'ambiguïté. Par exemple, les paires *Krais–Kreis* et *Krais–Kraus* obtiennent la même valeur BI-SIM, alors que la première paire, qui est correcte, est phonétiquement plus proche.

Nous gardons pour chaque mot w_{LND} les correspondances $(w_{\mathrm{LND}}, w_{\mathrm{LD}})$ qui maximisent la valeur BI-SIM, mais uniquement à condition que cette dernière se situe au-dessus du seuil de 0,7, choisi empiriquement. Ce seuil nous permet d'éliminer des correspondances qui sont très probablement fausses. Si plusieurs w_{LD} correspondent au même w_{LND}, nous les conservons tous. Il en résulte 10 369 mots néerlandais avec au moins une correspondance allemande, dont 7692 sont non-ambigus (c'est-à-dire qu'ils ont exactement un mot correspondant allemand). Le nombre de mots néerlandais distincts étant de 20 225, on a donc su trouver un correspondant en allemand dans 51,3% des cas. Pour la configuration palatin–allemand, nous obtenons 8437 mots palatins avec des correspondances allemandes (62,7%), dont 6568 non-ambigus.

3.1.2 Entraînement d'un modèle de traduction automatique

La prochaine étape consiste à utiliser les paires de cognats potentiels extraites comme indiqué ci-dessus pour entraîner un modèle de traduction automatique basé sur les caractères ; le but de cette étape est d'une part d'inférer des régularités dans les correspondances graphémiques entre les deux langues, et d'autre part d'obtenir un modèle plus robuste et moins sensible au bruit présent dans les données d'apprentissage que la seule mesure BI-SIM.

Les outils de traduction automatique basés sur GIZA++ et Moses disposent d'un nombre de paramètres élevé. Nous avons testé de nombreuses combinaisons de paramètres sur la paire néerlandais–allemand en fonction du taux d'exactitude sur le lexique d'évaluation présenté dans la section 3.1.4. Faute de place, nous ne pouvons pas décrire toutes ces expériences ici, et nous indiquons donc les paramètres tels que retenus *in fine*.

Modèle de langage Nous avons entraîné un modèle de langage sur le corpus Wikipedia de la langue dotée, c'est-à-dire l'allemand. Nous avons enlevé les mots apparaissant moins de 10 fois dans le corpus afin de limiter la taille et de réduire le bruit. Chaque mot est répété autant de fois qu'il apparaît dans le corpus. Le modèle à 10-grammes (de caractères) a

3. Ce seuil peut paraître très élevé, mais il constitue un bon compromis entre la qualité des résultats et la tractabilité informatique. S'il aurait effectivement été souhaitable d'utiliser le corpus entier, cela s'est avéré intractable puisque chaque mot de la langue source doit être comparé avec chaque mot de la langue cible.

4. Les expériences préalables ont montré qu'une configuration asymétrique, avec un peu plus de mots du côté de la LD, donnait de résultats légèrement supérieurs à une configuration purement symétrique.

 © ATALA

obtenu les meilleures performances. Nous avons utilisé IRSTLM (Federico *et al.*, 2008) pour l'entraînement du modèle de langage.

Combinaison d'alignements GIZA++ produit des alignements séparés dans les deux sens. Il existe différentes heuristiques pour combiner ces alignements et en extraire des segments. La combinaison *grow-diag-final* s'est montrée la plus efficace dans nos expériences.

Distortion Dans la terminologie de la TAS, la distortion se réfère à la possibilité de changer l'ordre des éléments. Elle peut être apprise de manière lexicalisée (favorisant le déplacement de certains mots au détriment d'autres), ou non (un poids unique). Nous avons préféré désactiver cette possibilité pour éviter d'apprendre des alignements croisés, que nous supposons très rares dans le contexte de correspondances de mots entre langues proches.

Lissage Nous utilisons le *Good Turing discounting* pour ajuster les poids des alignements rares.

Tuning Moses prévoit la possibilité d'adapter divers poids à l'aide d'un corpus de développement et du *Minimum Error Rate Training*. La taille limitée de nos données a conduit à des résultats peu satisfaisants avec cette approche. Nous avons donc gardé les poids par défaut.

3.1.3 Filtrage des candidats

Le modèle de TAS à base de caractères que nous utilisons permet de générer des candidats allemands qui n'ont jamais été vus tels quels par le modèle de langage. Si ce comportement est souhaitable par exemple pour recombiner des mots composés, il ne l'est pas lorsque les candidats proposés n'existent pas dans la langue allemande. [5] Une première idée pour éliminer ces mauvaises correspondances consistait à générer 50 candidats pour chaque mot de la LND et à en sélectionner le premier qui était un mot attesté dans la Wikipedia allemande. Ce filtre s'est avéré insuffisant : par exemple, le candidat erroné *Frankrijk* "France" était validé parce que ce mot néerlandais apparaissait 21 fois dans la Wikipedia allemande, à côté de 120 000 occurrences du candidat correct *Frankreich*. Nous avons donc choisi un filtre qui repose sur le produit entre le score obtenu par Moses et le score de similarité entre les fréquences des deux mots du candidat dans chacune des deux langues (« score combiné »).

De plus, nous avons constaté que beaucoup de paires induites avec un score faible sont fausses. Considérant qu'il vaut mieux n'obtenir aucune réponse qu'une réponse fausse, nous avons mis en place une stratégie de filtrage complémentaire. Elle consiste à éliminer tous les candidats dont le score combiné est inférieur à 0,5 écarts-type en-dessous de la moyenne de l'ensemble des scores combinés.

3.1.4 Évaluation des paires de cognats potentiels

Pour la paire néerlandais-allemand, nous avons procédé à une évaluation intermédiaire des différentes étapes d'induction lexicale présentées ci-dessus. Nous avons construit à cette fin un dictionnaire bilingue de 26 000 entrées, utilisé comme référence. Il est constitué de paires $(w_{\mathrm{LND}}, w_{\mathrm{LD}})$ extraites à partir des Wiktionnaires (Hanoka et Sagot, 2012) et de la collection OPUS (Tiedemann, 2012). Ces deux sources sont largement complémentaires : les entrées lexicales des Wiktionnaires sont précises, mais couvrent uniquement les formes de base. Pour

5. Notre système a par exemple proposé all. **Bienengarnelen* "crevettes d'abeilles" pour nl. *bijeengeroepen* "convoqué", ou all. **Knuffelfest* "fête de Knuffel" pour nl. *knuffelbeest* "peluche".

© ATALA

	BI-SIM (3.1.1)		Moses (3.1.2)		Filtre fréquence (3.1.3)		Filtre confiance (3.1.3)	
Total	20225		20225		20225		20225	
Sans réponse	9856	(48,7%)	0		0		4040	(20,0%)
Inconnu du lexique	4952	(24,5%)	11061	(54,7%)	11061	(54,7%)	8533	(42,2%)
Correct	2478	(12,3%)	3528	(17,4%)	3938	(19,5%)	3836	(19,0%)
Incorrect	2939	(14,5%)	5636	(27,9%)	5226	(25,8%)	3816	(18,9%)

TABLE 2 – Évaluation des différentes étapes à l'aide du lexique néerlandais

Mot néerlandais	Candidats cognats allemands			Commentaire
	retenus par BI-SIM (3.1.1)	meilleur résultat Moses (3.1.2)	retenus après filtrages (3.1.3)	
vegetatie	vegetative vegetation	vegetation	**vegetation**	désambiguïsation correcte par Moses
groenen	grossen großen groben	groenen	**grünen**	correct, grâce à Moses et au filtrage
amfibieën	—	amphibien	**amphibien**	correct mais inconnu de la référence
enkel	enkel onkel	enkel	**enkel**	faux amis (all. *einfach*)
zweden	zwecken zweigen	zweiten	**zweiten**	faux (all. *schweden*)

TABLE 3 – Exemples d'alignements obtenus pour le néerlandais, à différentes étapes du processus

également pouvoir évaluer les formes fléchies, nous avons choisi d'intégrer les entrées d'OPUS, bien que plus bruités. [6] La table 2 résume les taux de rappel et d'exactitude des différentes étapes, en partant des 20 225 mots-types de notre extrait de la Wikipedia néerlandaise.

Pour l'étape décrite en section 3.1.1, 48,7% des mots néerlandais ne sont associés à aucun candidat allemand dont la mesure BI-SIM dépasse le seuil de 0,7, et 24,5% ne peuvent être évalués directement car ils n'apparaissent pas dans le lexique de référence. Parmi ces derniers se trouvent principalement des noms propres, faciles à traduire puisque souvent identiques dans les deux langues, mais peu informatifs quant aux correspondances graphémiques régulières. Enfin, parmi les paires dont le mot néerlandais est connu, 45,7% (12,3% du total) sont correctes. Après utilisation de Moses (section 3.1.2), nous disposons désormais d'un candidat cognat allemand pour tous les mots néerlandais. Le nombre de paires évaluables passe de 5417 à 9164, sans que les scores de précision ne décroissent trop : 38,5% d'entre elles (soit 17,4% du total) sont correctes. Le filtrage par fréquence permet de faire passer la proportion de paires correctes parmi les paires évaluables à 43,0% (19,5% du total). Le filtrage par confiance réduit à 7652 le nombre de paires évaluables, mais avec une exactitude qui atteint 50,1% (3836 paires).

Quelques exemples concrets de cette évaluation sont donnés dans la table 3.

6. L'alignement des mots d'OPUS a été effectué uniquement à l'aide des propriétés distributionnelles, sans tenir compte des similarités orthographiques. Il est dès lors peu probable qu'une paire de mots, induite de manière erronée par notre algorithme de détection de cognats, ait également été induite par l'algorithme d'alignement d'OPUS.

© ATALA

3.1.5 Transfert d'étiquettes morphosyntaxiques

Les sections 3.1.1 à 3.1.3 ont décrit comment nous obtenons des paires de cognats $(w_{\text{LND}}, w_{\text{LD}})$. Il nous reste à créer un dictionnaire morphosyntaxique de la LD, c'est-à-dire un ensemble de couples de la forme (w_{LD}, c), où c est une étiquette. Dans notre cas, ce dictionnaire est extrait du corpus TIGER ; chaque mot allemand est associé de manière non-ambigüe avec son étiquette la plus fréquente. En fusionnant ces deux listes, nous pouvons finalement créer des paires de cognats catégorisés $(w_{\text{LND}}, w_{\text{LD}}, c)$ qui nous permettent d'annoter une partie du corpus.

Cette procédure peut échouer dans deux situations : premièrement, lorsqu'aucun candidat proposé par Moses passe le filtrage, et deuxièmement, lorsque le candidat proposé par Moses ne se trouve pas dans le dictionnaire morphosyntaxique. Dans ces cas, nous essayons de chercher le mot w_{LND} directement dans le dictionnaire morphosyntaxique de la LD, en comptant sur le fait qu'un certain nombre de mots ont une graphie identique dans les deux langues ; ceci permet notamment de récupérer les étiquettes des symboles de ponctuation, des abréviations, des nombres et des noms propres.

3.2 Extension du lexique des mots pleins par analyse de suffixes

Si un mot plein n'a toujours pas reçu de catégorie syntaxique à ce stade, nous essayons de la retrouver au moyen d'une analyse de suffixes. Nous identifions le plus long suffixe commun au mot non-identifié et à au moins un mot identifié, et nous associons au mot non-identifié l'étiquette associée au mot identifié. Au cas où plusieurs mots identifiés partagent le même suffixe, nous choisissons l'étiquette majoritaire de ces mots. Lorsqu'aucun suffixe du mot inconnu n'a été vu (c'est-à-dire qu'aucun mot ne se termine par le même caractère), nous étiquetons le mot comme *FM* (*foreign material*).

Pour ces deux heuristiques comme pour l'approche décrite au paragraphe suivant, des résultats quantitatifs seront donnés à la section 4.

3.3 Étiquetage des « mots grammaticaux » par similarité contextuelle

Comme indiqué précédemment, l'approche par cognats est peu appropriée pour les « mots grammaticaux ». Nous essayons donc d'inférer les étiquettes morphosyntaxiques pour ces mots à l'aide des contextes morphosyntaxiques immédiats. L'idée est la suivante : le corpus annoté allemand nous apprend que les mots de type *ART* (article) apparaissent souvent entre une préposition et un nom ou entre le début de la phrase et un nom. Si dans le corpus LND semi-annoté, le mot *de* a une distribution similaire, l'étiquette *ART* lui sera associée. La difficulté est que seule une partie du corpus LND est annotée, et ces annotations ne sont pas toujours correctes. Il s'agit donc de rendre cette procédure robuste aux données éparses et bruitées.

Plusieurs contextes syntaxiques peuvent être pertinents ; si le mot immédiatement précédent et le mot immédiatement suivant sont probablement les plus informatifs, on peut également considérer des contextes non-contigus. Nous récoltons les contextes suivants (où W dénote le mot dont on veut déterminer la catégorie, T dénote un mot dont on connaît la catégorie, et X dénote un mot dont on ne connaît pas la catégorie) : TWT, TWXT, TXWT, TW, WT, TXWXT, TWXXT, TXXWT. Considérons par exemple le mot néerlandais *is* "est" (verbe) dans la phrase partiellement

 © ATALA

	Moses	Mots identiques	Suffixe	*Foreign Material*	« mots grammaticaux »	**TOTAL**
Palatin — *évaluation sur 100 phrases de la Wikipedia annotées manuellement*						
Couverture	20,0%	18,1%	24,3%	0,9%	36,6%	100%
Précision	71,3%	92,7%	27,8%	16,7%	62,1%	**60,7%**
Néerlandais — *évaluation sur 26 022 phrases de la Wikipedia étiquetées en STTS par MElt (voir texte)*						
Couverture	20,6%	16,9%	19,9%	0,4%	42,2%	100%
Précision	59,2%	86,0%	48,4%	4,8%	73,0%	**67,2%**

TABLE 4 – Résultats de l'évaluation de l'annotation des corpus bruts pour le palatin et le néerlandais à partir des ressources pour l'allemand. Le pourcentage de mots traités par chaque approche est indiqué, ainsi que la précision de chaque approche sur les mots qu'elle a étiquetés.

étiquetée suivante : *het/?? is/?? waarschijnlijk/ADJD dat/?? de/?? trompetsignalen/NN van/?? de/?? vorige/ADJA dag/?? (...)*. Nous en extrayons entre autres le contexte *is ADJD*, de type WT, ou encore le contexte *[début de phrase] X is ADJD*, de type TXWT. Par contre, nous ne pouvons pas extraire de contexte de type TW, puisque l'étiquette du mot précédant *is* est inconnue.

Pour chaque type de contexte, par exemple TXWT, nous représentons alors chaque mot de la LND et chaque catégorie non-ouverte de la LD (ici, l'allemand) par un vecteur dont chaque élément correspond à un contexte particulier et en indique la fréquence : un contexte compte autant de fois qu'il existe de séquences de mots distincts qui lui correspondent. Pour chaque type de contexte et chaque mot grammatical, on classe les catégories candidates selon la similarité décroissante ; nous utilisons la distance euclidienne pour comparer les vecteurs. Les résultats obtenus pour chaque type de contexte sont ensuite moyennés à l'aide de la moyenne harmonique.

À l'issue de cette étape, tous les mots du corpus brut en LND ont été annotés. Il s'agit naturellement d'une annotation perfectible, que l'on pourrait par exemple utiliser comme corpus d'entraînement pour un étiqueteur probabiliste classique. Nous reviendrons sur ce point à la section 5 lorsque nous évoquerons quelques pistes pour des travaux futurs.

4 Résultats et discussion

L'objectif de ce travail est l'étiquetage morphosyntaxique de textes extraits de la Wikipedia palatine et de la Wikipedia néerlandaise, en s'appuyant uniquement sur la proximité étymologique de ces langues avec l'allemand. En plus de l'évaluation ponctuelle de l'induction des paires de cognats néerlandais–allemands (cf. section 3.1), nous présentons ici deux évaluations globales du modèle proposé, une pour la paire néerlandais–allemand, et une pour palatin–allemand.

Pour la première, nous utilisons comme référence la Wikipedia néerlandaise étiquetée par l'étiqueteur classique MElt (Denis et Sagot, 2012). Ce dernier a été entraîné sur le corpus Eindhoven (Uit den Boogaard, 1975; van Grootheest, 1992), dont nous avons converti les étiquettes au standard STTS au préalable. Pour la deuxième, nous utilisons comme référence un sous-ensemble de 100 phrases de la Wikipedia palatine que nous avons annoté manuellement. [7]

7. Cette annotation a été effectuée par le premier auteur, locuteur natif de l'allemand mais pas du dialecte palatin, en

 © ATALA

La table 4 montre les résultats de ces deux évaluations, et notamment le pourcentage de mots traités par chacune des approches décrites en section 3 (« Couverture »). On constate que la plus grande partie des mots-occurrences est traitée par l'approche des mots grammaticaux, à cause de la fréquence élevée de ces derniers. Les lignes intitulées « Précision » montrent, indépendemment pour chaque approche, le taux de précision. Les mots identiques obtiennent les meilleurs scores à cause des signes de ponctuation, qui sont très faciles à étiqueter. L'approche Moses et l'approche « mots grammaticaux » obtiennent des résultats comparables. En dernière colonne figure la précision totale de l'étiquetage. Si ces valeurs sont bien en deçà des résultats obtenus habituellement pour l'étiquetage morphosyntaxique, nous les considérons comme très satisfaisants vu la difficulté de la tâche : rappelons que nous ne disposons d'aucune information morphosyntaxique des LND et d'aucun alignement entre mots LND et mots LD.

5 Conclusion et perspectives

Nous avons proposé une méthodologie pour l'annotation en morphosyntaxe d'un corpus brut pour une langue peu dotée à partir de ressources pour une langue étymologiquement proche (lexique morphosyntaxique, corpus annoté, corpus brut volumineux). Cette tâche est plus complexe que la tâche classique d'étiquetage morphosyntaxique non-supervisé, pour laquelle on suppose disponible un lexique morphosyntaxique de la langue peu dotée. Nous avons appliqué cette technique aux paires de langues allemand–palatin et allemand–néerlandais, cette dernière étant destinée à permettre une évaluation plus systématique et un paramétrage de l'approche.

La méthode présentée ici est améliorable de plusieurs façons. Tout d'abord, nous prévoyons de chercher à introduire de l'ambiguïté de façon contrôlée, l'assignation d'une seule étiquette par mot n'étant clairement pas satisfaisante. C'est également une condition pour pouvoir mettre en œuvre de façon utile un étiqueteur morphosyntaxique, qui apprendrait à désambiguïser l'étiquetage en fonction des contextes d'occurrence. Il pourrait ainsi permettre l'étiquetage direct de nouveaux corpus dans la langue non dotée et faciliter l'extraction automatique du lexique morphologique extrait de tels corpus annotés. Par ailleurs, nous avons l'intention d'approfondir la distinction pour l'instant rudimentaire entre mots pleins et « mots grammaticaux ». Par exemple, le mot palatin *gheat* "appartient" est considéré comme un mot grammatical à cause de sa longueur et sa fréquence, et sa catégorie est induite de manière incorrecte. Mais s'il avait été considéré comme mot plein, Moses aurait inféré la correspondance correcte, et il aurait été étiqueté correctement. De plus, nous ne considérons actuellement pas la forme des « mots grammaticaux », bien qu'une partie non-négligeable de ces mots soient identiques dans les deux langues.

Enfin, nous souhaitons valider notre méthodologie sur d'autres paires de langues, qui impliquent d'autres langues dotées que l'allemand, si possible en variant les propriétés typologiques des langues concernées. Nous envisageons ainsi d'appliquer notre approche sur des paires de langues romanes, de langues slaves, mais également sur différents états d'une même langue, par exemple pour induire des ressources adaptées à des textes vieux de quelques siècles à partir de ressources adaptées aux textes contemporains.

Remerciements Ce travail a été financé par le LabEx EFL (ANR/CGI) dans le cadre de l'opération LR2.2.

moins d'une journée de travail.

 © ATALA

Références

BRANTS, S., DIPPER, S., HANSEN, S., LEZIUS, W. et SMITH, G. (2002). The TIGER Treebank. *In Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002)*, pages 24–41, Sozopol, Bulgarie.

BRILL, E. (1995). Unsupervised learning of disambiguation rules for part of speech tagging. *In Proceedings of the Third Workshop on Very Large Corpora*, pages 1–13, Cambridge, USA.

DENIS, P. et SAGOT, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, 46(4):721–736.

FEDERICO, M., BERTOLDI, N. et CETTOLO, M. (2008). IRSTLM : an open source toolkit for handling large scale language models. *In Proceedings of Interspeech 2008*, Brisbane, Australie.

FELDMAN, A., HANA, J. et BREW, C. (2006). A cross-language approach to rapid creation of new morpho-syntactically annotated resources. *In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 549–554.

FIŠER, D. et LJUBEŠIĆ, N. (2011). Bilingual lexicon extraction from comparable corpora for closely related languages. *In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'11)*, pages 125–131.

FUNG, P. (1998). A statistical view on bilingual lexicon extraction : from parallel corpora to non-parallel corpora. *Machine Translation and the Information Soup*, pages 1–17.

GOLDBERG, Y., ADLER, M. et ELHADAD, M. (2008). EM can find pretty good HMM POS-taggers (when given a good start). *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, pages 746–754, Columbus, USA.

GOLDWATER, S. et GRIFFITHS, T. L. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 744–751, Prague, République tchèque.

HANOKA, V. et SAGOT, B. (2012). Wordnet extension made simple : A multilingual lexicon-based approach using wiki resources. *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turquie.

HAUER, B. et KONDRAK, G. (2011). Clustering semantically equivalent words into cognate sets in multilingual lists. *In Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP'11)*, pages 865–873, Chiang Mai, Thaïlande.

INKPEN, D., FRUNZA, O. et KONDRAK, G. (2005). Automatic identification of cognates and false friends in French and English. *In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)*, pages 251–257.

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A. et HERBST, E. (2007). Moses : Open source toolkit for statistical machine translation. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07), demonstration session*, Prague, République tchèque.

KOEHN, P. et KNIGHT, K. (2002). Learning a translation lexicon from monolingual corpora. *In Proceedings of the ACL 2002 Workshop on Unsupervised Lexical Acquisition (SIGLEX 2002)*, pages 9–16, Philadelphia, PA.

 © ATALA

KOEHN, P., OCH, F. J. et MARCU, D. (2003). Statistical phrase-based translation. *In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT'03)*, pages 48–54.

KONDRAK, G. et DORR, B. (2004). Identification of confusable drug names : A new approach and evaluation methodology. *In In Proceedings of COLING 2004*, pages 952–958.

MANN, G. S. et YAROWSKY, D. (2001). Multipath translation lexicon induction via bridge languages. *In Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*, pages 151–158, Pittsburgh, PA, USA.

MERIALDO, B. (1994). Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2):155–72.

OCH, F. J. et NEY, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

RAPP, R. (1999). Automatic identification of word translations from unrelated English and German corpora. *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 519–526, Maryland, USA.

RAVI, S. et KNIGHT, K. (2009). Minimized models for unsupervised part-of-speech tagging. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP'09)*, pages 504–512, Singapour.

SMITH, N. et EISNER, J. (2005). Contrastive estimation : Training log-linear models on unlabeled data. *In Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 354–362, Ann Arbor, USA.

THIELEN, C., SCHILLER, A., TEUFEL, S. et STÖCKERT, C. (1999). Guidelines für das Tagging deutscher Textkorpora mit STTS. Rapport technique, Université de Stuttgart et Université de Tübingen.

TIEDEMANN, J. (2009). Character-based PSMT for closely related languages. *In Proceedings of the 13th Conference of the European Association for Machine Translation (EAMT 2009)*, pages 12 – 19, Barcelone.

TIEDEMANN, J. (2012). Parallel data, tools and interfaces in OPUS. *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turquie.

TIEDEMANN, J. et NABENDE, P. (2009). Translating transliterations. *International Journal of Computing and ICT Research*, 3(1):33–41. Special Issue of Selected Papers from the fifth international conference on computing and ICT Research (ICCIR 09), Kampala, Uganda.

UIT DEN BOOGAARD, P., éditeur (1975). *Woordfrequenties in geschreven en gesproken Nederlands*. Oosthoek, Scheltema en Holkema, Utrecht.

van GROOTHEEST, D. (1992). Handleiding bij het Eindhoven Corpus (VU-versie). Rapport technique, Vrije Universiteit Amsterdam.

VILAR, D., PETER, J.-T. et NEY, H. (2007). Can we translate letters ? *In Proceedings of the Second Workshop on Statistical Machine Translation*, pages 33–39, Prague, République tchèque.

YAROWSKY, D., NGAI, G. et WICENTOWSKI, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. *In Proceedings of the first international conference on Human language technology research*, San Diego, USA.

 © ATALA

Es esch fàscht wie Ditsch, oder net? *
Étiquetage morphosyntaxique de l'alsacien en passant par l'allemand

Delphine Bernhard[1] Anne-Laure Ligozat[2, 3]
(1) LiLPa, Université de Strasbourg
(2) LIMSI-CNRS, Orsay
(3) ENSIIE, Évry
dbernhard@unistra.fr, annlor@limsi.fr

RÉSUMÉ

L'analyse morphosyntaxique est une pierre angulaire de nombreuses applications du traitement automatique des langues. Elle nécessite toutefois des efforts de développement conséquents, qu'il s'agisse d'annoter des corpus ou de produire des lexiques et des outils. Pour les langues moins dotées, il peut être utile d'exploiter leur proximité avec des langues disposant d'outils et de ressources. Dans cet article, nous nous intéressons plus particulièrement aux dialectes alsaciens, qui présentent de nombreuses similitudes avec l'allemand standard. Nous montrons qu'il est possible d'utiliser des outils développés pour l'allemand afin de réaliser l'analyse morphosyntaxique de textes en alsacien. La méthode consiste à transposer les mots outils des textes alsaciens vers leurs équivalents en allemand standard. Cette transposition nécessite pour seule ressource un lexique bilingue des mots outils.

ABSTRACT

Es esch fàscht wie Ditsch, oder net ? POS-Tagging the Alsatian Dialects through German

Morphosyntactic analysis is a cornerstone of many natural language processing applications. However, it requires substantial development efforts, in order to annotate corpora or produce lexicons and tools. For under-resourced languages, it may be useful to exploit their proximity to languages for which tools and resources have already been developed. In this article, we focus on the Alsatian dialects, which have many similarities with standard German. We show that it is possible to use tools developed for the German language in order to perform the morphosyntactic analysis of Alsatian texts. The method consists in transposing function words in the Alsatian texts to their translations in standard German. This transposition requires as only resource a bilingual lexicon of function words.

MOTS-CLÉS : étiquetage morphosyntaxique, dialecte, alsacien, allemand.

KEYWORDS: POS tagging, dialect, Alsatian, German.

*. C'est presque comme de l'allemand, non ?

© ATALA

1 Introduction

Les dialectes alsaciens, parlés dans le nord est de la France, appartiennent aux familles des langues alémaniques et franciques (Huck *et al.*, 2007). Selon une étude récente, 43% de la population alsacienne se déclare dialectophone (OLCA / EDinstitut, 2012). Toutefois, la proportion des locuteurs de l'alsacien décroit régulièrement depuis les années 1960.

Le traitement automatique des dialectes alsaciens présente plusieurs difficultés :

– Il n'existe pas de standard graphique utilisé et adopté par tous (voir section 2) ;
– L'alsacien correspond à un continuum de dialectes, se caractérisant par des variantes lexicales et phonétiques ;
– Les ressources et corpus écrits sont rares.

Ces contraintes, qui se retrouvent pour nombre de dialectes et langues minoritaires, rendent difficile le développement d'outils de Traitement Automatique des Langues (TAL) selon les méthodes classiques. En effet, l'absence de corpus annotés rend impossible l'utilisation de méthodes par apprentissage supervisé. Par ailleurs, il est également difficile de trouver une main d'œuvre suffisante et qualifiée pour développer des ressources ou annoter des corpus.

Dans cet article, nous nous intéressons à l'une des premières étapes dans toute chaîne de traitement de données textuelles et présentons une méthode simple mais néanmoins efficace pour réaliser l'analyse morphosyntaxique de textes alsaciens. Nous proposons de transposer les mots grammaticaux (déterminants, pronoms, prépositions, conjonctions) et les auxiliaires dans leur équivalent en allemand standard puis d'utiliser des étiqueteurs morphosyntaxiques existant pour l'allemand. Cette approche permet d'exploiter la proximité de l'alsacien et de l'allemand standard.

L'article est structuré comme suit : la section 2 détaille la problématique de l'alsacien écrit. La section suivante décrit les travaux sur l'étiquetage morphosyntaxique des textes "hors norme". Nous décrivons notre approche dans la section 4 et présentons les résultats de l'évaluation dans la section 5.

2 L'alsacien écrit

L'alsacien appartient aux groupes alémaniques et franciques et se rapproche de fait des dialectes parlés dans les régions limitrophes d'Allemagne et de Suisse. On retrouve des emprunts au français, mais la morphosyntaxe est très similaire à celle de l'allemand standard.

La graphie des dialectes alsaciens n'est pas codifiée. L'OLCA (Office pour la Langue et la Culture de l'Alsace) précise d'ailleurs sur sa page Web de définition de la langue régionale : *«La langue normalisée, écrite et codifiée correspondante à nos dialectes est l'allemand standard.»* [1] Cette page cite également le recteur Pierre Deyon qui en 1985 affirmait qu'*«[i]l n'existe en effet qu'une seule définition scientifiquement correcte de la langue régionale en Alsace, ce sont les dialectes alsaciens dont l'expression écrite est l'allemand.»*

1. `http://www.olcalsace.org/fr/observer-et-veiller/definition-de-la-langue-regionale`.

Auteurs : Adrien Finck, Frédéric Hartweg, Raymond Matzen, Marthe Philip.

 © ATALA

Cette définition ne prend toutefois pas en compte les situations dans lesquelles il est nécessaire de passer par l'écrit pour documenter (dictionnaires, lexiques), transmettre (cours de langue), informer (encyclopédie, articles de presse, sites Web) et créer (littérature, pièces de théâtre). On trouve en effet de nombreux exemples d'alsacien écrit, notamment :

- Lexiques et dictionnaires : lexiques thématiques de l'OLCA [2], *De elsässisch Dico* (Bitsch et Matzen, 2004), *Mon premier dictionnaire français-alsacien en images* (Matzen *et al.*, 1997), etc. ;
- Manuels d'apprentissage de l'alsacien : *Wie geht's ?* (Matzen et Daul, 2004), *Parlons alsacien* (Schimpf et Muller, 1998), *L'alsacien pour les nuls* (Keck *et al.*, 2010), etc. ;
- Articles dans la presse régionale : L'Alsace et Dernières Nouvelles d'Alsace ;
- Sites Web d'entreprises locales, comme par exemple Wattwiller [3] ;
- Pièces de théâtre en alsacien : le théâtre dialectal est très vivace en Alsace et de nombreuses troupes perpétuent cette tradition dans les villages, ainsi que dans les agglomérations plus importantes (comme le Théâtre Alsacien de Strasbourg).

Il faut également souligner qu'il y a eu des tentatives pour définir une graphie propre à l'alsacien :

- Le système ORTHAL (Zeidler et Crévenat-Werner, 2008) se réfère à la norme graphique de l'allemand standard tout en permettant la transcription des spécificités des dialectes alsaciens. On retrouve notamment l'emploi des majuscules pour les substantifs et pronoms personnels à la forme de politesse, comme en allemand [4].
- Le système GRAPHAL-GERIPA (Hudlett et Groupe d'Etudes et de Recherches Interdisciplinaires sur le Plurilinguisme en Alsace et en Europe, 2003) spécifie un ensemble de règles pour passer des sons aux graphèmes.

Il est toutefois difficile d'estimer la diffusion et l'utilisation réelles de ces systèmes de scripturalisation. Par ailleurs, ils laissent naturellement place à la variation, afin de correspondre au mieux aux diverses variantes géolinguistiques rencontrées en Alsace. Ils ne peuvent donc résoudre à eux seuls le problème de la déviance par rapport à une certaine norme qu'il est nécessaire de respecter pour permettre un fonctionnement optimal des outils du TAL.

3 Étiquetage morphosyntaxique de textes "hors norme"

L'absence ou le non respect des conventions d'écriture est un problème récurrent en TAL car les outils sont généralement développés pour des données standard, exemptes de fautes d'orthographe ou de phénomènes déviants.

Giesbrecht et Evert (2009) évaluent des étiqueteurs morphosyntaxiques pour l'allemand pour des données issues du Web et montrent que leurs performances sont dégradées pour ce type de textes. Ils montrent également que la performance est dépendante du genre et que certains genres spécifiques du Web, comme les forums, sont plus difficiles à traiter.

Différentes approches ont été proposées pour contourner ces problèmes. La solution la plus fréquente consiste à normaliser les mots "hors norme". Mosquera *et al.* (2012) présentent

2. `http://www.olcalsace.org/fr/lexiques`
3. `http://www.wattwiller.com/?lang=de`
4. Il faut toutefois noter que tous les parlers alsaciens n'intègrent pas ces formes de politesse.

 © ATALA

une méthode permettant de transformer des textes du Web 2.0 vers leur forme canonique en anglais ou en espagnol. La méthode se décompose en deux étapes : dans un premier temps, les mots hors vocabulaire sont repérés, puis ces mots sont convertis vers leur forme normalisée. Cette conversion met en œuvre des mesures de similarité phonétique et graphique. Scherrer (2008) propose une approche à base de transducteurs stochastiques pour passer d'un dialecte de Suisse allemand (bernois) vers l'allemand standard, afin d'induire des lexiques bilingues. Hulden *et al.* (2011) visent à transformer des textes en dialecte Labourdin vers le basque standard. Ils utilisent pour cela des mécanismes d'apprentissage afin d'apprendre des transformations à partir d'un corpus parallèle. Deux méthodes sont proposées : la première extrait des règles phonologiques de remplacement en repérant les sous-chaines différentes dans les mots alignés d'un corpus parallèle ; la seconde est similaire à la première mais utilise également les contre-exemples pour chaque règle afin de déterminer les contextes les plus appropriés pour appliquer une règle de transformation. L'approche de Dasigi et Diab (2011) est légèrement différente puisqu'elle consiste à regrouper les variantes orthographiques d'un même lexème pour des dialectes de l'arabe standard. La méthode repose sur la classification non supervisée (*clustering*) et des mesures de similarité graphique (distance d'édition) et contextuelle (cosinus de vecteurs de co-occurrence).

Dans le contexte de l'étiquetage morphosyntaxique, il est possible d'entraîner un nouvel étiqueteur, adapté aux document à traiter. Cette approche a notamment été adoptée par Dipper (2011) qui décrit le processus d'entraînement du TreeTagger pour différentes versions d'un corpus de textes en moyen haut-allemand (mots normalisés ou non, dialecte). Cette approche n'est toutefois réalisable que si l'on dispose de textes préalablement annotés.

Lorsqu'au contraire les données annotées sont insuffisantes ou inexistantes, il est nécessaire d'avoir recours à des outils et ressources développés pour une langue apparentée et mieux dotée. Hana *et al.* (2011) décrivent un étiqueteur pour le vieux tchèque, basé sur deux stratégies. La première consiste à transformer un corpus annoté de tchèque moderne de manière à le rendre similaire à du vieux tchèque. Les transformations se basent sur un nombre limité de règles de changement graphique. Ce corpus transformé est alors utilisé pour entraîner un étiqueteur, qui sera appliqué à un corpus de vieux tchèque modernisé. Ce dernier corpus est à nouveau transformé sous sa forme originale en vieux tchèque, qui est alors annotée et permet donc d'entraîner un étiqueteur. La seconde stratégie consiste à utiliser une ressource morphosyntaxique pour le vieux tchèque, de manière à avoir une approximation des probabilités d'émission de l'étiqueteur. Une des idées sous-jacentes est qu'il suffit de produire cette ressource pour les mots les plus fréquents, qui auront l'impact le plus important sur l'apprentissage.

Nous reprenons certaines des idées proposées par Hana *et al.* : (i) nous transformons partiellement des textes alsaciens en allemand standard et (ii) nous appliquons ces transformations aux mots les plus fréquents, c'est-à-dire les mots grammaticaux. A la différence de Hana *et al.*, ces transformations ne peuvent pas être basées sur des changements de sons et de graphèmes dans le cas de l'alsacien. En effet, ces changements ne peuvent être prédits de manière systématique, comme le montre la Figure 1 qui représente les variantes graphiques que l'on peut trouver dans différents lexiques du Web pour deux lexèmes alsaciens. Ces deux exemples montrent l'extrême variabilité des scripturalisations : choix des voyelles (*u-i-é-ù-ü*, *a-e-i*, etc.), ajout / suppression de caractères (*g* ou *j* en fin de mots), alternances consonantiques (*ch-sch*, *d-t*).

© ATALA

Français	Allemand	Anglais	Variantes alsaciennes
cuisine	Küche	kitchen	Kuch, Kucha, Kische, Khésche, Kùch, Kücha, Kuche, Kiche, Kuchi
lundi	Montag	monday	Mondàà, Mantig, Mandig, Mondàà, Mondoe, Mondàj, Maandi, Mandi

FIGURE 1 – Variantes alsaciennes pour deux lexèmes trouvées dans divers lexiques sur le Web.

4 Approche

Nous avons collecté un corpus de cinq documents afin d'évaluer la performance d'étiqueteurs allemands pour des textes alsaciens :

- *Alsace* : un article tiré du journal local "L'Alsace", intitulé "Zìmlig beschta Frìnd" et datant du 18 février 2012 [5].
- *Dictionnaire multilingue* : phrases d'illustration tirées du *Dictionnaire comparatif multilingue Français - Allemand - Alsacien - Anglais* de Paul Adolf (2006). Ces phrases présentent la particularité d'être pour la plupart disponibles sous forme de diverses variantes graphiques. De plus, la traduction des phrases en allemand standard est également fournie (voir Figure 2).
- *Duttlenheim* : le résumé d'une pièce de théâtre jouée par la compagnie locale du village de Duttlenheim et datant de 2004 [6].
- *Hoflieferant* : une page de la pièce de théâtre "D'r Hoflieferant" de Gustave Stoskopf, datant de 1906.
- *Wikipedia* : un article de la Wikipédia alémanique, au sujet du musée alsacien de Strasbourg, récupérée le 30 octobre 2012 [7].

FR : Je n'ai pas assez d'argent	Ìch hà nìtt g'nüä Gald.
DE : Ich habe nicht genug Geld.	Ìch hàbb nìtt gnüej Gald.
EN : I don't have enough money.	Ìch hàbb nìtt genüej Gald.
	Ìch hàbb nìtt gnüej Gëld.
	Ìch hàbb nìtt genüej Gëld.

FIGURE 2 – Variantes graphiques de la même phrase dans le *Dictionnaire comparatif multilingue Français - Allemand - Alsacien - Anglais* de Paul Adolf (2006)

Ces textes ont été annotés de manière semi-automatique en utilisant un jeu simplifié d'étiquettes morphosyntaxiques afin d'obtenir une annotation de référence (voir Table 1). Nous avons tout d'abord traduit manuellement les mots en allemand approximatif puis avons corrigé les étiquettes fournies par le TreeTagger (Schmid, 1994) pour le *Dictionnaire multilingue* et le Stanford POS Tagger (Toutanova *et al.*, 2003) pour les autres textes (voir Figure 4). Une méthode d'analyse semi-automatique similaire a été utilisée par Giesbrecht et Evert (2009) pour obtenir leurs données de référence.

5. http://www.lalsace.fr/actualite/2012/02/18/zimlig-beschta-frind
6. http://theatreduttlenheim.free.fr/html/annee2004.htm
7. http://als.wikipedia.org/wiki/Els\%C3\%A4ssisches_Museum_\%28Stra\%C3\%9Fburg\%29

© ATALA

Les textes alsaciens ont été étiquetés par le TreeTagger et le Stanford POS Tagger en utilisant deux types de données :
- Textes originaux découpés manuellement en mots, sans autre pré-traitement ;
- Textes dont les mots grammaticaux et les auxiliaires ont été transposés dans leurs équivalents en allemand standard : *àwer* → *aber*, *fer* → *für*, *isch* → *ist*, etc. (voir Figure 3)

ALS	Brüchsch	kenn	Angscht	ze	han	for	mich	,	papa	.
TRANS	Brüchsch	keine	Angscht	zu	haben	für	mich	,	papa	.
DE-GLOSS	Brauchst	keine	Angst	zu	haben	für	mich	,	Vater	.
EN-GLOSS	Need	no	fear	to	have	for	me	,	dad	.
FR-GLOSS	As besoin	pas	peur	à	avoir	pour	moi	,	papa	.

FIGURE 3 – Exemple de transposition (TRANS) à partir de l'alsacien (ALS) et gloses en allemand (DE-GLOSS), en anglais (EN-GLOSS), et en français (FR-GLOSS).

Etiquette	**Description**	**Equivalence avec le jeu d'étiquettes du TreeTagger et du Stanford Tagger**
ADJ	Adjectif	ADJA, ADJD
ADP	Préposition / Postposition	APPO, APPR, APPRART, APZR
ADV	Adverbe	ADV
CARD	Nombre cardinal	CARD
CONJ	Conjonction	KOKOM, KON, KOUI, KOUS
DET	Déterminant	ART
FM	Mots en langue étrangère	FM
ITJ	Interjection	ITJ
N	Nom	NE, NN
PRN	Pronom	PAV, PROAV, PDAT, PDS, PIAT, PIDAT, PIS, PPER, PPOSAT, PPOSS, PRELAT, PRELS, PRF, PWAT, PWAV, PWAS, PWS
PRT	Particule	PTKANT, PTKNEG, PTKVZ, PTKZU, PTKA
V	Verbe	VAFIN, VAIMP, VAINF, VAPP, VMFIN, VMPP, VVFIN, VVIMP, VVINF, VVIZU, VVPP
$,	Virgule	$,
$.	Ponctuation de fin de phrase	$.
$(	Autre ponctuation	$(

TABLE 1 – Tableau d'équivalence des étiquettes morphosyntaxiques.

Le lexique utilisé pour la transposition contient 133 entrées. Ce lexique a été constitué manuellement, à partir du corpus. Les items ambigus, c'est-à-dire les mots grammaticaux pouvant avoir plusieurs traductions différentes, ont été éliminés de ce lexique. La Table 2 détaille le nombre de mots du corpus, ainsi que le nombre et le pourcentage de lemmes inconnus du TreeTagger, avant et après transposition.

© ATALA

Document	# tokens	# transpositions	# lemmes inconnus avant		# lemmes inconnus après	
Alsace	320	101	194	60,6%	104	32,5%
Dict-multi	1 180	353	794	67,3%	484	41,0%
Duttlenheim	166	37	60	36,1%	33	19,9%
Hoflieferant	230	39	74	32,2%	45	19,6%
Wikipedia	396	130	248	62,6%	139	35,1%

TABLE 2 – Statistiques du corpus.

5 Résultats

Pour l'évaluation, nous comparons les textes annotés de manière automatique avec nos annotations manuelles. Les étiquettes détaillées fournies par le TreeTagger et le Stanford Tagger sont automatiquement associées à l'étiquette équivalente dans notre jeu simplifié. Les résultats de l'évaluation sont détaillés dans les Tableaux 3 pour le TreeTagger et 4 pour le Stanford Tagger.

Corpus	Texte original	Texte original, mots non transposés uniquement	Texte transposé	Texte transposé, mots non transposés uniquement
Alsace	0,48	0,69	0,79	0,74
Dict. multi.	0,53	0,61	0,79	0,79
Duttlenheim	0,67	0,78	0,86	0,83
Hoflieferant	0,64	0,73	0,78	0,76
Wikipedia	0,50	0,71	0,80	0,75

TABLE 3 – Précision du TreeTagger

Corpus	Texte original	Texte original, mots non transposés uniquement	Texte transposé	Texte transposé, mots non transposés uniquement
Alsace	0,56	0,74	0,86	0,83
Dict. multi	0,62	0,69	0,87	0,87
Duttlenheim	0,77	0,85	0,89	0,88
Hoflieferant	0,68	0,76	0,83	0,80
Wikipedia	0,53	0,70	0,85	0,80

TABLE 4 – Précision du Stanford Tagger.

La précision du TreeTagger pour les données originales est plutôt basse et se situe sous la barre de 0,70. Après la transposition, la précision augmente, avec un minimum de 0,78. Nous avons également calculé la précision uniquement pour les mots non transposés, afin de voir si la performance de l'étiquetage augmente uniquement pour les mots transposés ou s'il y a

 © ATALA

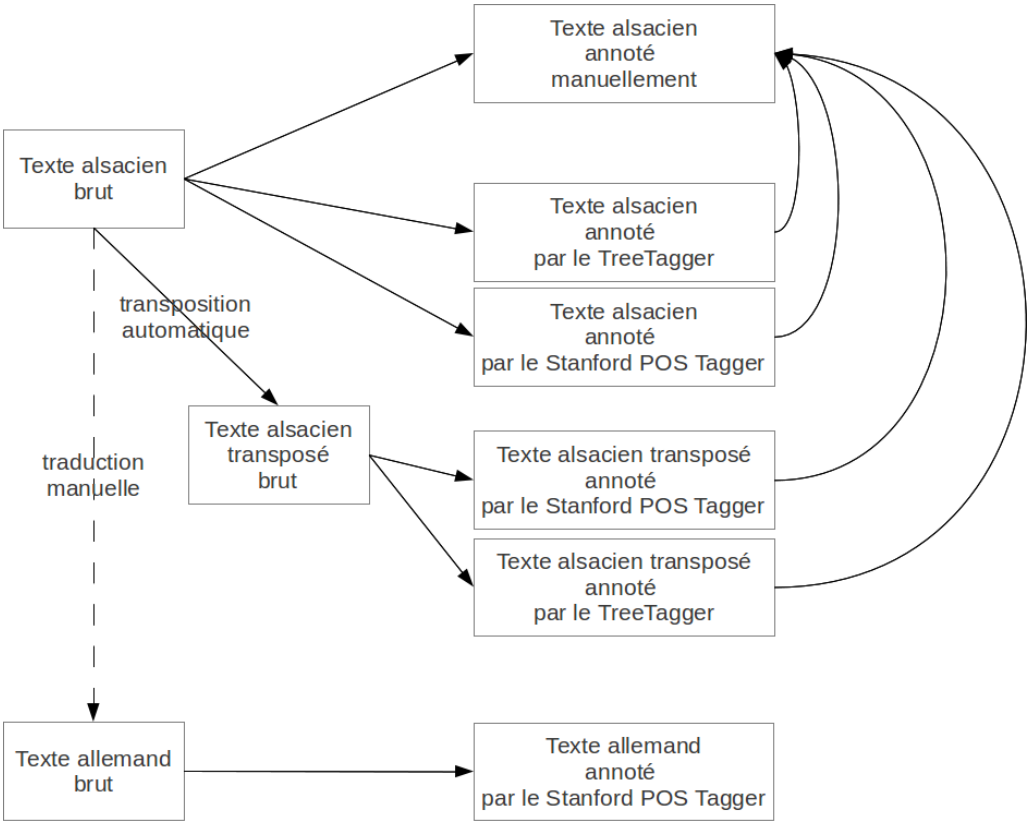


FIGURE 4 – Étiquetage morphosyntaxique de textes alsacien, avec ou sans transposition.

également une incidence sur les autres mots. Nous constatons que les mots non transposés sont également mieux analysés : par exemple, la précision augmente de 0,71 à 0,75 pour les mots non transposés du texte issu de Wikipédia.

Le Stanford Tagger a des résultats légèrement supérieurs à ceux du TreeTagger, probablement en raison du modèle sous-jacent. Comme pour le TreeTagger, les résultats sont meilleurs pour les textes transposés : par exemple, la précision augmente de 0,53 à 0,85 pour les textes tirés de Wikipédia. La précision pour les mots non transposés est également améliorée.

Nous avons analysé les résultats de manière plus détaillée, en nous focalisant sur certaines différences entre l'alsacien et l'allemand.

Absence de génitif L'allemand standard marque la possession en utilisant soit le génitif saxon, soit les compléments du nom introduits par la préposition "von" (de) suivie du datif. En alsacien, c'est cette dernière construction de datif possessif qui est privilégiée. On trouve ainsi "Wien ìsch d'*Hauptstàdt vun Eschtrich*." [8] en alsacien, alors qu'en allemand standard on

8. Vienne est la capitale de l'Autriche.

© ATALA

peut avoir "Wien ist *Österreichs Hauptstadt*." (génitif saxon) ou "Wien ist die *Hauptstadt von Österreich*." (source : *Dictionnaire multilingue*). Nous avons repéré six exemples de ce type dans notre corpus. Dans tous les cas, la construction est bien analysée. Cette construction se retrouvant en allemand standard, on peut comprendre qu'elle ne pose aucune difficulté aux analyseurs.

Formes périphrastiques avec l'auxiliaire "düen" Cet auxiliaire est couramment utilisé pour le présent et le conditionnel et fonctionne de manière similaire à l'auxiliaire "do" an anglais. Les formes périphrastiques alternent au présent avec les formes simples du verbe. Ces constructions ont été décrites de manière détaillée par Kleiber et Riegel (1998, 2005). Pour reprendre un des exemples de Kleiber et Riegel (2005, p. 171), on peut trouver la forme périphrastique "Gewiss, e Màmme düet sich fer ihri Kinder ufopfere" [9] ou la forme simple, sans l'auxiliaire "düen" : "Gewiss, e Màmme opfert sich fer ihri Kinder uf." La construction périphrastique ne se retrouve pas en allemand standard. Nous n'avons trouvé que deux exemples de cette construction au présent dans notre corpus, aucune au conditionnel. Si l'auxiliaire "düen" est dans tous les cas analysé comme un verbe, à la fois par le TreeTagger et le Stanford Tagger, l'analyse échoue pour le verbe principal qui est rejeté en fin de proposition. Cette construction est a priori inconnue des analyseurs entraînés pour l'allemand standard, ce qui peut expliquer les problèmes d'étiquetage pour le reste de la phrase. Il faudrait toutefois disposer d'un corpus plus conséquent pour pouvoir tirer des conclusions plus précises.

Utilisation du passé composé à la place de l'imparfait et du passé simple On ne trouve qu'un temps pour exprimer le passé en alsacien : le passé composé ou parfait (Schimpf et Muller, 1998; Keck *et al.*, 2010). Nous avons repéré huit exemples de cette construction dans notre corpus. Dans la majorité des cas, l'auxiliaire et le participe sont correctement analysés : en effet, le parfait est un temps qui existe également en allemand standard et qui peut donc être reconnu par les outils.

Omission du pronom personnel sujet "du" à la deuxième personne du singulier Ce pronom est fréquemment omis dans les dialogues à l'oral (voir Figure 3). Nous avons trouvé quatre exemples d'omission du pronom "du" dans deux documents du corpus. Dans tous les cas, les deux analyseurs ont donné une étiquette erronée au verbe : nom, adjectif ou mot en langue étrangère.

Phénomènes de code-switching On observe régulièrement des incursions du français en alsacien. Par exemple, dans l'extrait de la pièce de théâtre "D'r Hoflieferant" de notre corpus, on trouve la réplique suivante : "Toi, tais-toi !". Cette pratique est toutefois essentiellement liée à l'oral et ce phénomène est donc fréquent dans les médias alsaciens comme la télévision régionale (Erhart, 2012). Le code-switching peut concerner différents types d'unités : mots, syntagmes, phrases complètes. Dans notre corpus, nous avons trouvé quelques exemples de code switching uniquement dans l'extrait de la pièce de théâtre "D'r Hoflieferant" ("Eh bien", "Toi, tais-toi !"). Dans les autres textes, le français n'est utilisé que pour nommer des entités difficiles à traduire (film, association, livre). Le problème se pose toutefois dès l'annotation

9. Sûr, une maman se sacrifie pour ses enfants.

© ATALA

manuelle : comment étiqueter ce type de phénomènes ? Faut-il donner l'étiquette appropriée en français, ou alors l'étiquette FM (Foreign Material : mots en langue étrangère) ? C'est cette dernière solution que nous avons privilégiée lors de l'annotation.

L'analyse des principales différences entre l'alsacien et l'allemand standard tend à montrer qu'il faudrait prétraiter certaines constructions, notamment les formes périphrastiques avec l'auxiliaire "düen" et l'absence de pronom personnel sujet à la deuxième personne du singulier, avant analyse par les étiqueteurs. En particulier, l'absence de pronom personnel sujet devrait pouvoir être résolue par l'insertion du pronom avant étiquetage. Les formes périphrastiques avec l'auxiliaire "düen" constituent un autre exemple de difficulté, mais dont la résolution semble plus difficile, en raison du déplacement du verbe principal en fin de proposition.

6 Conclusion et perspectives

L'analyse morphosyntaxique est une étape importante des applications de TAL Nous avons présenté une méthode pour l'analyse morphosyntaxique des dialectes alsaciens qui repose sur une transposition partielle vers l'allemand standard. Les résultats montrent qu'une approche simple qui consiste à transposer uniquement les mots grammaticaux conduit à des performances d'étiquetage améliorées.

Il reste toutefois une grande marge de progression, de manière à atteindre des niveaux de précision au delà des 95%, correspondant aux niveaux de performance attendus pour la tâche.

Une première piste que nous souhaitons explorer est celle de l'identification des cognats en alsacien et en allemand. En effet, les mots pleins alsaciens ont souvent une forte ressemblance graphique avec leurs équivalents en allemand standard. Ces cognats pourraient ensuite être transposés de manière automatique et améliorer ainsi l'étiquetage des mots pleins.

Une autre possibilité consisterait à intégrer les informations sur la catégorie disponible dans des lexiques alsaciens existants, disponibles sur le Web. Ces informations pourraient permettre de corriger les erreurs de l'analyse automatique. Il faudra toutefois dans ce cas pouvoir résoudre les problèmes de variation graphique car les mots des documents ne se trouvent pas forcément avec la même graphie dans des lexiques rédigés par des auteurs différents.

Enfin, le prétraitement de certaines difficultés syntaxiques identifiées pourrait également conduire à une amélioration des résultats.

Remerciements

Nous remercions Paul Adolf pour nous avoir fourni une version numérique de son dictionnaire multilingue, Chafik Taleb pour la transformation du dictionnaire au format XML et Pascale Erhart pour sa relecture. Ces travaux ont bénéficié du soutien du conseil scientifique de l'université de Strasbourg dans le cadre du projet COPAL (COrpus Parallèles pour l'ALsacien).

 © ATALA

Références

ADOLF, P. (2006). *Dictionnaire comparatif multilingue : français-allemand-alsacien-anglais*. Midgard, Strasbourg, France.

BITSCH, R. et MATZEN, R. (2004). De elsässisch Dico. CD, Editions l'Ami hebdo, Strasbourg et site Web : `http://www.ami-hebdo.com/elsadico/index.php`.

DASIGI, P. et DIAB, M. (2011). CODACT : Towards Identifying Orthographic Variants in Dialectal Arabic. *In Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 318–326, Chiang Mai, Thailand.

DIPPER, S. (2011). Morphological and Part-of-Speech Tagging of Historical Language Data : A Comparison. *Journal for Language Technology and Computational Linguistics*, 26(2):25–37.

ERHART, P. (2012). *Les dialectes dans les médias : quelle image de l'Alsace véhiculent-ils dans les émissions de la télévision régionale ?* Thèse de doctorat, Université de Strasbourg.

GIESBRECHT, E. et EVERT, S. (2009). Is Part-of-Speech Tagging a Solved Task ? An Evaluation of POS Taggers for the German Web as Corpus. *In Proceedings of the 5th Web as Corpus Workshop (WAC5)*.

HANA, J., FELDMAN, A. et AHARODNIK, K. (2011). A low-budget tagger for Old Czech. *In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH '11)*, pages 10–18.

HUCK, D., BOTHOREL-WITZ, A. et GEIGER-JAILLET, A. (2007). L'Alsace et ses langues. Éléments de description d'une situation sociolinguistique en zone frontalière. *Aspects of Multilingualism in European Border Regions : Insights and Views of Alsace, Eastern Macedonia and Thrace, the Lublin Voivodeship and South Tyrol*, page 13–100.

HUDLETT, A. et GROUPE D'ETUDES ET DE RECHERCHES INTERDISCIPLINAIRES SUR LE PLURILINGUISME EN ALSACE ET EN EUROPE (2003). *Charte de la graphie harmonisée des parlers alsaciens : système graphique GRAPHAL - GERIPA*. Centre de Recherche sur l'Europe littéraire (C.R.E.L.), Mulhouse, France.

HULDEN, M., ALEGRIA, I., ETXEBERRIA, I. et MARITXALAR, M. (2011). Learning word-level dialectal variation as phonological replacement rules using a limited parallel corpus. *In Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, page 39–48.

KECK, B., DAUL, L. et KRETZ, P. (2010). *L'alsacien pour les nuls*. Pour les nuls (Éd. de poche), ISSN 1625-0486. Paris, France.

KLEIBER, G. et RIEGEL, M. (1998). Grammaticalisation et auxiliaire modal : L'énigme de düen en alsacien. *In Travaux de linguistique*, volume 36, pages 161–173.

KLEIBER, G. et RIEGEL, M. (2005). Les périphrases düen + verbe à l'infinitif en alsacien : Un auxiliaire modal à tout... faire. *In Les Périphrases Verbales*, volume 25 de *Linguisticae Investigationes Supplementa*, pages 171–184. John Benjamins Publishing Company.

MATZEN, R. et DAUL, L. (2004). *Wie geht's ? : le dialecte à la portée de tous*. La Nuée bleue, Strasbourg, France.

MATZEN, R., DAUL, L. et LAZÉ, C. (1997). *Mon premier dictionnaire français-alsacien en images*. Gisserot J.P, Paris, France.

 © ATALA

MOSQUERA, A., LLORET, E. et MOREDA, P. (2012). Towards Facilitating the Accessibility of Web 2.0 Texts through Text Normalisation. *In Proceedings of the Worshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*.

OLCA / EDINSTITUT (2012). Étude sur le dialecte alsacien. [En ligne, visité le 7 mai 2013 : `http://www.olcalsace.org/sites/default/files/documents/etude_linguistique_olca_edinstitut.pdf`].

SCHERRER, Y. (2008). Transducteurs à fenêtre glissante pour l'induction lexicale. *In Actes de RECITAL 2008*, Avignon.

SCHIMPF, J.-P. et MULLER, R. (1998). *Parlons alsacien*. L'Harmattan, Paris, France.

SCHMID, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *In Proceedings of the International Conference on New Methods in Language Processing*, page 44–49.

TOUTANOVA, K., KLEIN, D., MANNING, C. D. et SINGER, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, page 173–180.

ZEIDLER, E. et CRÉVENAT-WERNER, D. (2008). *Orthographe alsacienne : bien écrire l'alsacien de Wissembourg à Ferrette*. J. Do Bentzinger, Colmar, France.

© ATALA

Jochre, océrisation par apprentissage automatique : étude comparée sur le yiddish et l'occitan

Assaf Urieli Marianne Vergez-Couret[1]
CLLE-ERSS, UMR5263, Université de Toulouse 2 Le Mirail
assaf.urieli@univ-tlse2.fr, marianne.vergez@univ-tlse2.fr

RÉSUMÉ

Pour constituer des bases de textes pour des langues peu dotées comme le yiddish et l'occitan, il faut créer des outils et des ressources permettant une reconnaissance optique de caractères (océrisation) de qualité. Une des difficultés principales à surmonter pour ces deux langues est leur grande variation graphique (et dialectale pour l'occitan). Il est généralement admis qu'un lexique augmente la qualité de l'océrisation, mais la façon dont il faut tenir compte de la variation au sein du lexique n'est pas toujours clair. Dans cette étude, nous avons utilisé un logiciel d'apprentissage automatique supervisé, Jochre. Nous comparons plusieurs façons de constituer et d'utiliser les lexiques. La meilleure méthode nous permet d'atteindre une exactitude de 91,2 % (mots) et 97,4 % (lettres) sur le corpus yiddish, et de 93,2 % (mots) et 97,9 % (lettres) pour le corpus occitan.

ABSTRACT

Supervised machine learning OCR using Jochre : a comparative study of Yiddish and Occitan

To create textual databases for less-resourced languages such as Yiddish and Occitan, we need tools and resources enabling high-quality OCR (optical character recognition). One of the main difficulties to overcome for these two languages is their considerable spelling variation (and dialectal variation for Occitan). It is generally admitted that a lexicon can improve OCR quality, but it is not clear how to take such variation into account within the lexicon. In this study, we use Jochre, a supervised machine learning OCR system. We compare several methods of generating and using lexicons. The best method allows us to attain an accuracy of 91.2% (words) and 97.4% (letters) for the Yiddish corpus, and 93.2% (words) and 97.9% (letters) for the Occitan corpus.

MOTS-CLÉS : OCR, océrisation, yiddish, occitan, lexique, dialectes, apprentissage automatique.
KEYWORDS : OCR, Yiddish, Occitan, lexicon, dialects, supervised machine learning.

1 Introduction

Dans cette étude, nous nous intéressons à l'océrisation, à l'aide du logiciel Jochre, de deux langues peu dotées en ressources TAL : le yiddish et l'occitan, que nous présentons dans les sous-sections 1.1 et 1.2. Nous étudions plus particulièrement l'apport du lexique dans le contexte d'un système d'apprentissage automatique supervisé.

Il est souvent dit que l'océrisation des textes imprimés est un problème résolu, ce qui est vrai pour des documents récents en anglais, mais l'exactitude baisse énormément dès qu'il s'agit de langues peu dotées en ressources informatiques, ou de documents dont la qualité des

[1] Avec le soutien du Conseil Régional de Midi-Pyrénées et l'université de Toulouse Le Mirail dans le cadre d'un financement de post-doctorat.

 © ATALA

images est dégradée. Cette exactitude peut se mesurer de deux façons : au niveau des mots et au niveau des lettres. Tanner, Muñoz et Ros (2009) mesurent une exactitude pour les lettres de 99,98 % pour des journaux anglais édités après 1950, de 95 % pour ceux édités entre 1900 et 1950, et de < 85 % pour ceux édités avant 1900. Les études récentes sont rares pour d'autres langues. On peut citer Boschetti et al. (2009) qui atteignent une exactitude pour les lettres oscillant entre 92 % et 95 % pour le grec ancien dans des éditions critiques historiques.

Quelques systèmes libres d'OCR ont vu le jour récemment, dont Tesseract (Smith, 2007), ainsi que OCROpus (Breuel, 2008) et Jochre[2], qui mettent l'accent sur des techniques d'apprentissage automatique supervisé. Pour cette étude, nous avons choisi Jochre car il présentait plusieurs avantages par rapport aux autres logiciels libres : il intégrait déjà un modèle linguistique pour le yiddish, un module d'intégration de lexiques, et un module d'évaluation sophistiquée. Ce logiciel sera présenté en section 2.

Les techniques d'apprentissage supervisé nécessitent la constitution de corpus annotés pour l'entraînement, et parallèlement de lexiques pour l'évaluation. La constitution de ces ressources est décrite dans la section 3, et les résultats de notre étude dans la section 4.

1.1 Occitan

1.1.1 La langue occitane

L'occitan est une langue romane, parlée dans le sud de la France, dans le Val d'Aran en Espagne et dans 12 vallées alpines d'Italie. Faute d'étude spécifique, il est difficile d'estimer le nombre de locuteurs occitanophones. Diverses études annoncent des chiffres allant de 600 mille à 2 millions de locuteurs (Martel, 2007). Cette langue connaît une variation interne relativement importante. Elle regroupe un ensemble de variétés organisées en dialectes :

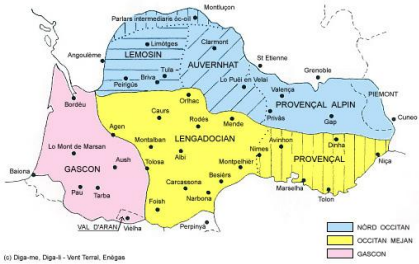


FIGURE 1 – Dialectes de l'occitan (Carles, 2005)

Dans cet article, nous nous focalisons sur deux dialectes de la langue occitane : le languedocien et le gascon dont voici quelques exemples de variation : lo filh/eth hilh ; luna/lua et cabra/craba (Bec, 1995).

L'occitan n'est pas une langue unifiée et standardisée. Mais elle est écrite depuis le moyen-âge et a produit une littérature très importante. L'occitan s'écrit alors dans une graphie dite "des troubadours". Cette graphie disparaît peu à peu que la production littéraire décroit. Au XIXème siècle, une première graphie normalisée voit le jour en Provence, la graphie dite

[2] https://github.com/urieli/jochre

© ATALA

mistralienne. Au XXème siècle apparaît une orthographe unifiée, dite graphie classique inspirée de la graphie des troubadours (Sibille, 2007). Dans le cadre de cette étude, nous avons souhaité utiliser uniquement des œuvres écrites ou transcrites en graphie classique pour nous focaliser sur la variation dialectale. Toutefois, nous notons tout de même des variations d'un auteur à un autre, par exemple, on trouve dans un œuvre gasconne en graphie classique des vestiges de la graphie dite béarnaise (tabé vs. tanben ; coundes vs. condes). Il demeure également des variations liées à la normalisation qui ne s'est pas faite en un jour, par exemple sur les conjugaisons (avian vs. avián) ou des variations de prononciation (contes vs. condes).

1.1.2 Projet BaTelÒc

Le projet BaTelÒc[3] (Bras, 2006 ; Bras et Thomas, 2011) vise la construction d'une base de textes en langue occitane (sur le modèle de Frantext) en rassemblant des œuvres écrites de tous genres (littérature, théâtre, contes, textes techniques et journalistiques…) des périodes modernes et contemporaines. Un million de mots a déjà été rassemblé à partir d'œuvres contemporaines existant au format numérique. Suffisamment de matières est disponible pour envisager de passer à plusieurs centaines de millions de mots. Mais une partie de cette matière n'existe pas encore au format numérique et une étape préalable de numérisation et d'océrisation est alors nécessaire. Nous ne connaissons pas à ce jour de logiciel d'océrisation capable de traiter convenablement la variation graphique et dialectale autrement qu'en rentrant manuellement des listes de mots. Nous avons donc souhaité entraîner un modèle pour l'occitan et tester l'impact de l'utilisation de lexiques de formes fléchies construits de façon hiérarchique pour la langue, les dialectes et les parlers[4] sur les performances de l'océrisation. Le TALÒc (Traitement automatique de la langue occitane) fait intervenir une dimension nouvelle par rapport au TAL des langues très dotées : la gestion de la variation dialectale et graphique.

1.2 Yiddish

1.2.1 La langue yiddish

Le yiddish est à l'origine la langue des communautés juives ashkénazes d'Europe centrale et de l'est. Au début du 20ème siècle, on estime le nombre de yiddishophones de 11 à 13 millions (Jacobs, 2005). Suite au génocide nazi, à la russification forcée de l'époque stalinienne, à la position anti-yiddish du mouvement sioniste, et à l'assimilation des juifs aux Etats-Unis, ce nombre s'est réduit de nos jours à quelques centaines de milliers, dont la plupart font partie des communautés « hassidiques », pourtant en pleine croissance démographique (Katz, 2004). Le yiddish est une langue hybride, avec un vocabulaire de base tiré de l'allemand médiéval, de l'hébreu biblique, de l'araméen, et des langues slaves des pays où vivaient les juifs, mais souvent classé comme langue germanique. Il s'écrit avec l'alphabet hébreu, mais avec certaines différences majeures : les voyelles, largement absentes dans l'hébreu en dehors des signes diacritiques, sont réintroduites dans le yiddish en tant que lettres à part entière, sauf pour les mots d'origine hébraïque, qui sont écrits comme dans les sources religieuses, sans les voyelles.

[3] Projet dirigé par Myriam Bras (CLLE-ERSS Université de Toulouse 2) depuis 2006.
[4] Le terme parler est utilisé pour désigner une variante d'un dialecte sur une aire géographique plus restreinte (un village, une vallée...).

© ATALA

A l'instar de l'occitan, le yiddish est une langue sans gouvernement ni académie. Il y a deux normes orthographiques officielles : celle du YIVO[5] et celle du gouvernement soviétique. Le YIVO a cherché à définir un yiddish moderne et standard qui engloberait tous les dialectes, alors que le gouvernement soviétique a cherché à effacer tout rappel de l'hébreu dans le yiddish afin de l'éloigner de toute connotation religieuse. Cependant, jusqu'à 1950, la plupart des éditeurs en dehors de l'Union Soviétique ont appliqué leurs propres normes orthographiques, parfois inspirées de l'allemand (redoublement des consonnes à l'intérieur des mots, ajout de la lettre hébraïque « hey » pour imiter un H muet), et rarement cohérentes. Au niveau dialectal, les éditeurs se voulaient presque tous « universels », et les différences dialectales, très présentes à l'oral, sont largement absentes à l'écrit.

1.2.2 Le Corpus du Yiddish Book Center

Le Yiddish Book Center à Amherst, Massachussetts a constitué un corpus à partir des livres en yiddish retrouvés partout dans le monde (Lansky, 2004). Ils ont amassé 1,5 million de livres à ce jour, dont une grande partie a été distribuée à plus de 600 bibliothèques, universitaires et autres. Les livres comprennent 18 000 titres uniques, dont 12 000 ont été scannés, et peuvent être téléchargés gratuitement en ligne[6]. Ces livres sont en grande majorité édités entre 1870 et 1960, période qui correspond à la renaissance de la culture et de la littérature yiddish. Notre corpus d'entraînement est tiré entièrement de ce corpus.

2 Le logiciel OCR Jochre

Le logiciel Jochre[7] (*Java Optical CHaracter REcogition*) est un logiciel OCR libre développé par Assaf Urieli, au départ pour océriser le corpus du Yiddish Book Center avec des techniques d'apprentissage automatique supervisé. Pour cette étude, nous l'avons adapté à l'occitan. L'analyse de Jochre s'effectue en trois étapes : 1) segmentation des images en paragraphes, lignes, groupes et formes ; 2) reconnaissance des lettres et 3) correction des mots à l'aide du lexique.

2.1 Segmentation

La segmentation utilise des techniques statistiques ad hoc adaptées à chaque tâche (détection de l'orientation, suppression des petites taches, …). A la sortie de la segmentation, on a une *forme* pour chaque lettre, et un *groupe* de formes pour chaque mot. Dans cet article, un *groupe* signifie donc « séquence de formes graphiques qui correspond à un seul mot dans la page ». Il arrive que Jochre ne soit pas capable de trouver la segmentation exacte d'une image. En particulier, deux lettres peuvent être fusionnées en une seule forme (trop d'encre lors de l'impression, tâche), et une lettre peut être scindée en deux formes (manque d'encre, livre abîmé). Les taux d'erreur de segmentation sont de 2,6 % pour le corpus yiddish, et de 12,5 % pour le corpus occitan. Cette étude ne cherche pas à améliorer la segmentation de Jochre : quand deux lettres sont fusionnées, Jochre va tenter de trouver les deux lettres à la fois, et quand une lettre est scindée, il va tenter de trouver les deux moitiés de la lettre

[5] YIVO : le Yidisher Visnshaftlekher Institut, ou Institut Scientifique Yiddish, établie à Vilnius en 1925 en tant que référence dans l'étude de la langue yiddish.
[6] http://www.yiddishbookcenter.org/books/search
[7] https://github.com/urieli/jochre

© ATALA

séparément lors de la reconnaissance des lettres.

2.2 Reconnaissance des lettres

La reconnaissance des lettres s'applique à chaque forme retrouvée lors de l'étape de segmentation. Elle utilise des techniques d'apprentissage automatique, et se divise donc en deux activités distinctes : l'entraînement et l'analyse.

Lors de l'étape d'*entraînement*, on sélectionne un corpus d'apprentissage pour la langue qu'on souhaite océriser en scannant des documents (quelques pages par document). Les images scannées sont ensuite chargées dans le logiciel web JochreTrain qui effectue automatiquement la segmentation de ces images (étape précédente). A travers l'interface JochreTrain, l'utilisateur attribue manuellement la bonne lettre à chaque forme. Ensuite, il faut choisir les *traits* qui caractérisent les formes. Un trait est n'importe quelle information concernant la forme qui peut aider le logiciel à choisir la bonne lettre. Par exemple, on peut imaginer un trait PetitPointEnHaut qui donne « vrai » si la forme contient un petit point en haut, et « faux » dans le cas contraire. Pour le français, un résultat « vrai » pour ce trait peut indiquer que la forme est un « i » à 95 % ou un « j » à 5 %. La dernière étape de l'entraînement est la construction du modèle statistique. Pour ceci, on utilise un *classifieur*, qui est le « moteur » chargé de décider de l'importance relative à accorder à chaque trait. Il va ainsi attribuer un poids à chaque trait et à chaque lettre, en cherchant à maximiser la vraisemblance de ces poids par rapport au corpus d'apprentissage. La grande matrice des poids par trait et par lettre est ce qu'on appelle le *modèle statistique*. Dans les expériences décrites ici, on utilise un classifieur par entropie maximale, basé sur Ratnaparkhi (1998).

Lors de l'étape d'*analyse*, l'utilisateur fournit à Jochre une nouvelle image à océriser. Jochre va segmenter l'image, appliquer les traits à chaque forme trouvée dans l'image segmentée, et appliquer le modèle statistique aux résultats des traits pour générer une distribution de probabilités des lettres pour cette forme (selon les probabilités observés dans le corpus).

2.2.1 Traits de base (baseline)

Pour la reconnaissance des lettres, nous avons utilisé une liste de traits de base, qui s'appliquent à n'importe quel alphabet et langue. Parmi ceux-ci se trouvent les traits *n-grammes*, qui indiquent les séquences de lettres les plus probables dans une langue donnée. Le *n* réfère au nombre des lettres dans la séquence. En français, par exemple, la séquence « pro » est bien plus probable que la séquence « pry ». Donc, si pendant l'analyse on a déjà choisi « p » pour la première forme et « r » pour la deuxième, le trait 3-grammes nous indique que la troisième lettre est plus probablement un « o » qu'un « y ».

La liste complète des traits de base est la suivante :

- Ngram2 : la lettre reconnue par Jochre pour la forme précédente
- Ngram3 : les lettres reconnues par Jochre pour les deux formes précédentes
- SectionRelativeBrightnessGrid : La noirceur relative de chaque section d'une grille (divisée en 9 sections verticales et 5 sections horizontales) par rapport à la noirceur de la section la plus sombre.
- VerticalSize : La hauteur verticale de la forme, divisée par l'œil typographique

© ATALA

(hauteur d'une lettre minuscule comme « x » sans jambage supérieur ni inférieur).

- VerticalElongation : Le rapport entre la hauteur et la largeur de la forme.
- BaselineDistance : La limite inférieure de la forme par rapport à la ligne de pied.
- LastShapeInSequence : Est-ce que cette forme est la dernière dans le groupe actuel ?
- TouchesBaseLine : Est-ce que cette forme touche la ligne de pied ?
- TouchesMeanLine : Est-ce que cette forme touche la ligne en haut de l'œil ?

2.2.2 Traits spécialisés pour l'alphabet hébreu

Pour le yiddish, nous avons aussi développé des traits spécifiques aux paires de lettres qui sont souvent confondues par le logiciel quand on applique uniquement les traits de base.

FIGURE 2 – Les lettres similaires *daled* et *reysh* dans l'alphabet hébreu

Dans la FIGURE 2 on voit deux lettres hébraïques très similaires, *daled* et *reysh*. On a donc développé un trait qui cherche une protubérance dans le coin en haut à droite de la lettre. Des traits semblables ont été développés pour distinguer 9 paires de lettres. Le développement de ces traits « spécialisés » est assez coûteux en temps. Du coup, pour l'occitan, nous avons basé notre évaluation uniquement sur les traits de base.

2.3 Correction à l'aide d'un lexique

Lors de la reconnaissance des lettres, Jochre applique une technique de « recherche par faisceau » (*beam search* ou *breadth-first search*)) (Bisiani, 1992). Pour chaque groupe, cette technique va fournir les *n* séquences de lettres les plus probables, construites à partir des distributions des probabilités des lettres pour chaque forme dont le groupe est constitué. Il s'agit maintenant de choisir le bon mot parmi ces séquences. En particulier, Jochre peut utiliser un lexique pour modifier les scores de chacune des *n* séquences. Un *lexique* ici est une liste de formes lexicales fléchies possibles dans une langue donnée, avec éventuellement une fréquence associée à chaque forme. Si la séquence est inconnue dans le lexique, on va multiplier son score par un coefficient de réduction < 1.

acordat	score initial	connu ?	score ajusté
acordot	**72,0 %**	non (x 0,5)	36,0 %
acordat	70,1 %	**oui** (x 1,0)	**70,1 %**
acordet	64,3 %	non (x 0,5)	32,2 %

TABLE 1 – Analyse avec un coefficient de réduction de 0,5 et un faisceau de 3

Le TABLE 1 ci-dessus montre l'analyse du mot « acordat » par Jochre, avec un lexique qui contient ce mot, un coefficient de réduction des mots inconnus de 0,5, et une largeur de faisceau de 3. A la sortie de la reconnaissance des lettres, Jochre propose donc les 3 séquences de lettres les plus probables : « acordot » (72,0 %), « acordat » (70,1 %), et

© ATALA

« acordet » (64,3 %). Noter que ces scores se calculent en prenant la moyenne harmonique des probabilités pour chaque lettre de la séquence. Sans lexique, c'est « acordot » qui aurait été choisi. Avec le lexique, les scores pour les mots inconnus sont multipliés par un coefficient de 0,5, et c'est « acordat » qui se trouve en tête. Il est aussi possible de privilégier les mots fréquents, si le lexique indique la fréquence relative des mots. Dans ce cas, Jochre nous permet d'utiliser la formule : $\text{Score}_{\text{final}} = \text{Score}_{\text{initial}} \times (1 + \log_k(\text{freq}))$

où $Score_{initial}$ est le score fourni par la fonction de reconnaissance des lettres pour la séquence en question, k la base du logarithme, fourni comme paramètre à Jochre, et $freq$ la fréquence du mot indiqué par le lexique. Plus k est élevée, moins la fréquence va compter dans le score final. Par exemple, pour k=10, le score d'un mot dont la fréquence est 1 restera inchangé, le score d'un mot dont la fréquence est 10 sera multiplié par 2, et le score d'un mot dont la fréquence est 100 sera multiplié par 3.

3 Corpus annoté et lexiques

3.1 Préparation des corpus

Les corpus d'entraînement et d'évaluation sont constitués à partir de plusieurs pages (>60) de livres variés de sorte à diversifier les types, les tailles et les styles (gras, italique…) de police (environ 6 pages par livre). Pour l'occitan, le corpus est également diversifié du point de vue dialectal : 39 pages pour le gascon et 41 pages pour le languedocien.

	Corpus Yiddish	**Corpus Occitan**
Nombre de livres numérisés	14	10[8]
Années d'édition	1910-1955	1960-2000
Lieu d'édition	Europe, Amérique du Nord, Amérique du Sud	France
Nombre de pages	95	80
Nombre de mots	27 400	20 400
Nombre de lettres	123 400	85 500

TABLE 2 – Caractéristiques des corpus

Les corpus ont été corrigés après une première analyse automatique du corpus en les confrontant là où l'annotation manuelle et l'annotation automatique diffèrent : l'annotation automatique permet dans ce cas de corriger des erreurs humaines et d'améliorer la qualité du corpus d'entraînement.

3.2 Préparation des lexiques de formes fléchies

3.2.1 Yiddish

Pour le Yiddish, nous avons utilisé les entrées en yiddish du dictionnaire bilingue yiddish-français de Yitskhok Niborski et Bernard Vaisbrot (Niborski et Vaisbrot, 2002). Ce

[8] 4 auteurs présents dans le corpus, Rouquette, Blader, Laux et Mouly, ont un statut spécial car nous disposons pour ces auteurs de lexiques extraits de leurs œuvres, cf. section 3.2.2.

© ATALA

dictionnaire, avec 37 000 mots, a la couverture lexicale la plus large de tous les dictionnaires yiddish bilingues parus jusqu'à présent. La version informatique de la partie yiddish du dictionnaire nous a été fournie par Harry Bochner, éditeur de la traduction en anglais de ce dictionnaire (Beinfeld et Bochner, 2013). Nous avons développé un logiciel pour transformer cette liste automatiquement en une liste de 396 500 formes fléchies. Le dictionnaire de Niborski et Vaisbrot contient très peu de noms propres. Pour combler ce manque, nous avons compilé, avec l'aide du Yiddish Book Center, des listes de prénoms et diminutifs tirées au départ de Harkavy (1928), des noms de famille et des noms géographiques.

3.2.2 Occitan

L'occitan est encore à ce jour une langue peu dotée en matière de ressources lexicales et informatisées même s'il existe de nombreux dictionnaires sous format papier (Bras et Thomas, 2007). Pour commencer, nous avons tiré profit des textes déjà présents dans BaTelÒc pour construire des lexiques de formes fléchies pour le languedocien et le gascon en considérant toutes les formes trouvées dans ces œuvres.

	Languedocien	**Gascon**
Total (sans les doublons)	70400 formes	10100 formes
Dont les auteurs[9]	Laux (7700 formes)	Blader (5 300 formes)
	Mouly (9600 formes)	
	Rouquette (11 000 formes)	

TABLE 3 - Lexiques de formes fléchies par dialecte issus de BaTelÒc

Nous avons ensuite constitué des lexiques à partir des entrées de plusieurs dictionnaires numérisés : pour le gascon, le dictionnaire français/occitan du gascon toulousain (Rei Bèthvéder, 2004) qui contient des noms communs et des noms propres et pour le languedocien, le dictionnaire français/occitan (Laux, 2005) et le lexique occitan/français des mots occitans de Max Rouquette (Rouquette, 2010). Lorsque les informations de flexion étaient signalées, nous avons automatiquement construit les féminins des noms et des adjectifs. Enfin, nous avons construit les pluriels pour tous les noms et les adjectifs.

	Languedocien	**Gascon**
Dictionnaire/Lexique	Dictionnaire Laux (81 200 formes)	Dictionnaire Gascon Toulousain (21 900 formes)
	Lexique de G. Rouquette (3 800 formes)	
Total (sans les doublons)	84 376 formes	21 900 formes

TABLE 4 - Lexiques de formes fléchies par dialecte issus de dictionnaires et lexiques

Tous les lexiques ainsi constitués contiennent des informations de fréquence : la fréquence absolue pour les formes fléchies extraites des textes de BaTelÒc et une fréquence de 1 pour les entrées et leur(s) flexion(s) des dictionnaires ou lexiques bilingues.

[9] Uniquement les auteurs qui sont également présents dans le corpus.

© ATALA

Enfin, nous avons fusionné les lexiques de la façon suivante afin de tester pour les deux sous-corpus gascon et languedocien l'apport de chaque :

- Lex_Global (150 700 formes) : tous les lexiques réunis (tous dialectes confondus).
- Lex_Languedocien (135 300 formes): tous les lexiques en dialecte languedocien.
- Lex_Gascon (28 900 formes) : tous les lexiques en dialecte gascon.

Nous avons également fusionné des lexiques par auteur afin de tester l'impact de la précision de ces lexiques (représentant précisément le parler et la graphie de l'auteur) pour chaque sous-corpus correspondant à l'auteur.

- Lex_Rouquette (17 700 formes) : les lexiques de formes fléchies des textes de BaTelÒc et les formes issues du lexique des mots occitans de Rouquette.
- Lex_Laux (84 800 formes): les lexiques de formes fléchies des textes de BaTelÒc et les formes fléchies issues du dictionnaire du même auteur.
- Lex_Molin (9 600 formes): les lexiques de formes fléchies des textes de BaTelÒc.
- Lex_Blader (5 300 formes) : les lexiques de formes fléchies des textes de BaTelÒc.

4 Résultats

Pour chaque corpus, nous avons effectué une validation croisée en le divisant en six parts égales : à chaque fois nous avons entraîné sur 5/6èmes et évalué sur le 1/6ème restant, chaque 1/6ème étant tour à tour évalué. Les scores indiqués ici sont les scores totaux des 6 validations (nombre d'erreurs/nombre total). Ce score total est toujours très proche du score moyen des 6 validations avec un écart type inférieur à 1 %.

4.1 Yiddish

Nous avons mesuré les taux d'erreur 1) avec et sans lexique, et 2) avec et sans traits spécialisés pour le corpus entier et pour chaque œuvre séparément. Les meilleurs résultats sont de 91,20 % pour les mots et 97,44 % pour les lettres. Si on prend la mesure sans lexique et sans traits spécialisés comme baseline, ceci représente un gain de 31,7 % (mots) et de 30,6 % (lettres) par rapport à toutes les erreurs qu'il restait à corriger. Il y a une variation conséquente d'exactitude par œuvre, allant, pour les mots, de 78,93 % pour un livre de poésie alternant une police italique et une police décorative, à 97,78 % pour un livre relativement récent et bien conservé dans une police standard de type « Times ».

	Traits de base		**Traits spécialisés**	
	Mots	**Lettres**	**Mots**	**Lettres**
Sans lexique	87,11	96,31	89,74	97,09
Avec lexique	89,14	96.84	**91,20**	**97,44**

TABLE 5 - Taux de réussite par méthode pour le yiddish

Comme indiqué dans la TABLE 5, les traits spécialisés apportent un peu plus que le lexique, mais les meilleurs résultats sont obtenus par la combinaison des deux. Les traits spécialisés

© ATALA

permettent de mieux identifier les lettres au départ, alors que le lexique ne peut s'appliquer qu'aux *n* analyses les plus probables déjà identifiées.

4.2 Occitan

Pour l'occitan, nous avons mesuré les taux d'erreur avec et sans lexique. Les meilleurs résultats pour la totalité du corpus sont de 93,15 % pour les mots et 97,93 % pour les lettres. On note, comme pour le yiddish, une variation par œuvre allant, pour les mots, de 88,21 % pour une œuvre en police italique à 97,21 % pour une œuvre en police standard. Si on prend la mesure sans lexique comme baseline, ceci représente un gain de 19 % (mots) et de 16,2 % (lettres) par rapport à toutes les erreurs qu'il restait à corriger.

	Ensemble du corpus	
	Mots	**Lettres**
Sans lexique	91,54	97,53
Lex_Gascon	92,72	97,81
Lex_Languedocien	92,83	97,86
Lex_Global	**93,13**	**97,93**

TABLE 6 – Taux de réussite pour le corpus occitan total

Les résultats montrent un apport systématique des lexiques pour l'occitan, que ce soit un lexique global ou un lexique par dialecte. Les meilleurs résultats pour l'ensemble du corpus sont obtenus avec le lexique englobant languedocien et gascon. Nous avons, dans un second temps, comparé pour chaque sous-corpus gascon et languedocien l'impact des lexiques languedocien et gascon.

	Corpus Languedocien		**Corpus Gascon**	
	Mots	**Lettres**	**Mots**	**Lettres**
Sans lexique	92,08	97,64	90,99	97,41
Lex_Gascon	93,07	97,85	**92,36**	**97,78**
Lex_Languedocien	**94,10**	**98,15**	91,53	97,56
Lex_Global	94,08	98,13	92,16	97,71

TABLE 7 – Taux de réussite pour le corpus occitan par dialecte

Pour le sous-corpus languedocien, les meilleurs résultats sont de 94,10 % pour les mots et de 98,15 % pour les lettres (gain de 25,5 % (mots) et de 21,6 % (lettres) par rapport à toutes les erreurs qu'il restait à corriger) avec le lexique en languedocien mais ces résultats sont tout juste sensiblement meilleurs à ceux obtenus avec le lexique global (qui de fait est majoritairement composé de formes en languedocien dû au déséquilibre de la présence des deux dialectes dans le lexique). Pour le sous-corpus gascon, les meilleurs résultats sont de 92,36 % pour les mots et 97,78 % pour les lettres (gain de 15,2 % (mots) et 14,2 % (lettres) par rapport à toutes les erreurs qu'il restait à corriger) avec le lexique gascon. L'utilisation d'un lexique permet toujours d'apporter un gain, même en utilisant le lexique gascon sur le

© ATALA

corpus languedocien et le lexique languedocien sur le corpus gascon. En revanche, le lexique languedocien permet un gain de 25,5 % (mots) sur le corpus languedocien contre 6 % (mots) sur le corpus gascon. Inversement, le lexique gascon permet un gain de 15,2% (mots) sur le corpus gascon contre 12,5 % (mots) sur le corpus languedocien. Cela encourage prioritairement la création de lexiques de taille importante et en second lieu de lexiques par dialecte.

Enfin, nous regardons spécifiquement les sous-corpus des auteurs Blader (pour le gascon), Molin, Laux et Rouquette (pour le languedocien). Pour chacun des sous-corpus, nous cherchons le lexique qui a permis d'obtenir les meilleurs résultats pour les mots.

	Corpus_Blader	**Corpus_Laux**	**Corpus_Molin**	**Corpus_Rouquette**
Lex_Global	**97,62**	94,52	96,05	89,56
Lex_Languedocien	96,75	94,25	**96,1**	89,67
Lex_Gascon	97,53	93,89	95,44	87,11
Lexique de l'auteur	97,39	**94,34**	95,95	**89,95**

TABLE 9 – Pourcentage de réussite (mots) pour le corpus occitan par auteur

De bons résultats sont obtenus en utilisant le lexique global pour le sous-corpus en gascon (Corpus_Blader) et le lexique languedocien pour les trois autres sous-corpus. L'utilisation de petit lexique spécialisé par auteur permet une augmentation sensible des résultats pour le Corpus_Laux et le Corpus_Rouquette bien qu'ils soient jusqu'à 10 fois plus petits. Ces résultats encouragent, lorsque c'est possible, l'utilisation d'un lexique spécifique à l'auteur qui grossit à mesure que son œuvre est numérisée.

4.3 Autres paramètres

Concernant le coefficient de réduction du score pour les mots inconnus (*cf.* section 2.3.), nous avons testé des coefficients variants de 0,1 à 1,0, où un coefficient de 1,0 met les mots connus et inconnus au même niveau (comme s'il n'y avait pas de lexique). La FIGURE 3 montre l'exactitude des mots (connus, inconnus et les deux confondus) selon ce coefficient. Plus le coefficient augmente, plus l'exactitude des mots connus baisse, et celle des mots inconnus augmente. Comme il y a 27,0 % de mots inconnus dans le corpus yiddish et 14,9 % dans le corpus occitan, l'apport des mots connus est plus important. L'équilibre total est atteint dans les deux cas avec un coefficient de 0,75, ce qui est assez surprenant, vu la différence significative dans le mode de construction des deux lexiques. Ce qui est surtout étonnant dans ce graphique, c'est la différence importante qui subsiste entre les mots connus et les mots inconnus dans les deux corpus quand on atteint un coefficient de 1,0. Il faudrait analyser les données pour émettre des hypothèses quant à cette différence[10].

Nous n'avons pas testé l'apport de la fréquence pour le yiddish, car nos lexiques pour le yiddish sont tous construit à partir de dictionnaires. Pour l'occitan, prendre la fréquence en compte fait baisser les scores systématiquement. Plus on accord de l'importance à la fréquence, plus le score baisse. Même en utilisant un logarithme à base 1000 (c'est-à-dire

[10] Nous remercions le relecteur anonyme pour son hypothèse que nous pourrons tester ultérieurement, à savoir que les mots inconnus sont peut-être des mots plus longs.

© ATALA

que si un mot avait une fréquence de 1000, son score serait deux fois plus élevé que s'il avait une fréquence de 1), l'exactitude baisse de presque 1 %.

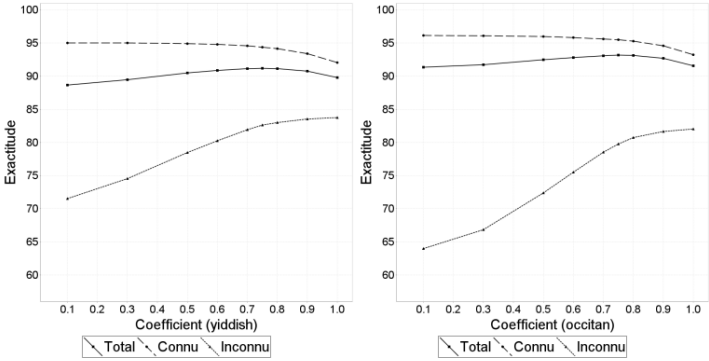


FIGURE 3 - Exactitude des mots selon le coefficient de réduction des mots inconnus

Concernant la recherche par faisceau, nous avons testé des faisceaux de 1, 2, 5, 10, 20 et 50. Un faisceau de 10 donne des résultats supérieurs aux autres faisceaux testés, ce qui voudrait dire qu'avec des faisceaux plus larges, le lexique choisit de façon erronée des mots peu probables, alors qu'avec des faisceaux moins larges, le lexique n'a pas toujours accès au bon mot dans la liste des séquences qui lui est fournie. Pour connaître l'apport maximal possible pour un lexique, il faut savoir, parmi les erreurs, combien de faisceaux contiennent la bonne réponse. Pour les erreurs en yiddish, avant d'appliquer le lexique, le faisceau de 2 contient 40 % des bonnes réponses, 5 contient 59 %, 10 contient 65 %, 20 contient 68 %, et 50 contient 69 %. Après l'application du lexique actuel au faisceau de 10, parmi les erreurs qui reste, le faisceau contenait la bonne réponse dans 59 % des cas, ce qui correspond à l'apport maximal possible pour un lexique à plus grande couverture. Pour l'occitan avec un faisceau de 10, ce pourcentage passe de 69 % sans lexique à 62 % avec le lexique actuel.

Le dernier point important à noter est la différence entre les mots bien segmentés et ceux qui contiennent au moins une lettre mal segmentée. Dans le corpus yiddish, les mots mal segmentés ne représentent que 2,6 % du corpus, mais leur taux d'erreur est de 58,7 % contre 7,5 % pour les autres. Dans le corpus occitan, les mots mal segmentés représentent 12,5 % du corpus, avec un taux d'erreur de 26,8 % contre 4,0 % pour les autres.

5 Conclusion et perspectives

Dans cette étude, nous nous sommes intéressés à l'océrisation, à l'aide du logiciel Jochre, de deux langues peu dotées en ressources TAL : le yiddish et l'occitan. Pour le yiddish, nous avons confirmé l'apport du lexique et des traits spécifiques à l'alphabet hébreu, mais la question reste ouverte sur la couverture plus faible du lexique comparé à l'occitan, alors que ce lexique est tiré d'un dictionnaire à large couverture, augmenté de listes de noms propres. Pour l'occitan, nous souhaitions évaluer l'apport des lexiques en faisant jouer d'une part les dialectes et d'autre part les parlers de plusieurs auteurs pour lesquels nous disposions de lexiques et d'œuvres numérisées. Les résultats montrent en premier lieu l'importance de disposer de lexiques de taille importante. Néanmoins, certains des résultats tendent à indiquer l'importance de la précision des lexiques, au niveau du dialecte, mais également au

© ATALA

niveau du parler spécifique de l'auteur. Il sera intéressant de regarder qualitativement là où le lexique apporte un gain ou induit le système en erreur (ce qui est le risque majeur de l'utilisation d'un lexique non approprié au dialecte ou au parler). De plus, seule cette analyse qualitative permettra de mesurer l'influence de la variation graphique *vs.* dialectale sur les performances de l'OCR.

Pour les deux langues, nos résultats mettent en avant deux classes de mots présentant une exactitude nettement plus basse : les mots mal segmentés et les mots inconnus dans le lexique. Pour les mots mal segmentés, outre améliorer les algorithmes ad hoc de segmentation, on pourrait imaginer un système d'apprentissage de la segmentation, entraîné sur le corpus d'apprentissage déjà constitué. Ce système tenterait de reconnaître des formes à scinder ou à fusionner, en étudiant les traits graphiques des formes déjà annotées en tant que lettre fusionnée ou scindée. Pour les mots inconnus, il faudrait procéder à une typologie, afin de cerner l'effort lexical qui serait le plus utile pour ajouter ce type de mot au lexique, sachant qu'il reste encore 59 % d'erreurs pour le yiddish et 62 % pour l'occitan qui peuvent être corrigées par un lexique à plus large couverture. Il faudrait également expliquer pourquoi une différence importante subsiste entre mots connus et inconnus même quand le lexique n'est pas utilisé : y-a-t-il un type de mot inconnu qui est intrinsèquement difficile à océriser et si oui, pourquoi ?

Au niveau du développement informatique, on a pu constater l'apport majeur des traits spécialisés à l'océrisation de l'alphabet hébreu. Il serait intéressant de développer des traits spécialisés pour l'alphabet latin, dans l'espoir de voir un apport semblable pour l'occitan.

Finalement, il nous semble intéressant d'explorer ce que l'océrisation peut apporter au lexicographe, dans l'optique d'enrichir les dictionnaires existants. Pendant l'analyse d'une œuvre, Jochre peut sortir une liste des formes inconnues dans le lexique, dont des erreurs d'analyse, des noms propres, des graphies inhabituelles, et des formes qui manquent dans les dictionnaires actuels. Ces dernières sont particulièrement intéressantes pour le lexicographe, d'autant plus qu'elles se trouvent dans le contexte d'une phrase tirée souvent d'une œuvre littéraire.

Remerciements

Pour le Yiddish, nous tenons à remercier Aaron Lansky, Catherine Madsen, Katie Palmer Finn, Joshua Price, Agnieszka Ilwicka du Yiddish Book Center pour leur aide dans la préparation du corpus et du lexique ; Yitskhok Niborski (INALCO) et Gilles Rozier (Bibliothèque Medem) pour leur accord concernant l'utilisation du dictionnaire de Niborski et Vaisbrot ; Harry Bochner pour avoir fourni ce dictionnaire en format XML ; et Paul Glasser du YIVO pour avoir fourni une liste des noms géographiques.

Pour l'occitan, nous tenons à remercier N. Rei Bèthvéder, C. Laux et G. Rouquette, l'IDECO et l'IEO du Tarn qui ont mis à disposition des versions électroniques d'œuvres, de dictionnaires et de lexiques pour la recherche et Myriam Bras pour les textes du projet BaTelÒc.

Références

BEC, P. (1995). *La langue occitane*. Que sais-je n°1059. Paris.

© ATALA

BEINFELD, S., et BOCHNER, H. (Eds.) (2013). *Comprehensive Yiddish-English Dictionary*. Indiana University Press.

BISIANI, R. (1992). Beam search. *Encyclopedia of Artificial Intelligence*, pages 1467-1468. Wiley-Interscience, 2nd edition. Editor: S. C. Shapiro.

BOSCHETTI, F., ROMANELLO, M., BABEU, A., BAMMAN, D., ET CRANE, G. (2009). Improving OCR accuracy for classical critical editions. *In Research and Advanced Technology for Digital Libraries*, pp. 156-167. Springer Berlin Heidelberg.

BRAS, M. (2006). Le projet TELOC : construction d'une base textuelle occitane. *In Langues et Cité : bulletin de l'observation des pratiques linguistiques,* 8, p9.

BRAS, M. et THOMAS, J. (2007). Diccionaris, corpora, e basas de donadas textualas, *In Linguistica Occitana, 5,* p. 1-22.

BRAS, M. et THOMAS, J. (2011). Batelòc : cap a una basa informatisada de tèxtes occitans. *In* A. Rieger (ed.) *L'Occitanie invitée de l'Euregio. Liège 1981 - Aix-la-Chapelle 2008 Bilan et perspectives*, Actes du IXème Congrès International de l'AIEO, Aache, Shaker.

BREUEL, T. M. (2008). The OCRopus open source OCR system. *Proceedings IS&T/SPIE 20th Annual Symposium.*

CARLES, S. (2005). *Diga-me, diga-li*, Vent Terral.

HARKAVY, A. (1928). *Yiddish-English-Hebrew Dictionary.* Hebrew Publishing Co., New York.

JACOBS, N. (2005) *Yiddish: a Linguistic Introduction*, Cambridge University Press.

KATZ, D. (2004). *Words on Fire: The Unfinished Story of Yiddish* , New York: Basic Books.

LANSKY, A. (2004). *Outwitting History: The Amazing Adventures of a Man Who Rescued a Million Yiddish Books.* Chapel Hill: Algonquin Books of Chapel Hill.

LAUX, C. (2005). *Dictionnaire Français-Occitan*. IEO del Tarn.

MARTEL, P. (2007). L'occitan, qui parle ? *In Langues et Cité : bulletin de l'observation des pratiques linguistiques, 10,* p3.

NIBORSKI, Y. et VAISBROT B. (2002). *Dictionnaire yiddish–français.* Bibliothèque Medem, Paris.

RATNAPARKHI, A. (1998) *Maximum entropy models for natural language ambiguity resolution*, PhD Thesis, University of Pennsylvania.

REI BETHVEDER, N. (2004). *Dictionnaire Français/Occitan Gascon Toulousain*. IEO edicions.

ROUQUETTE, J.-G. (2010). *Lexique Occitan-Français de Max Rouquette*. Association Amistats Max Rouquette.

SIBILLE, J. (2007). L'occitan, qu'es aquò ? In *Langues et Cité : bulletin de l'observation des pratiques linguistiques, 10*, p2.

SMITH, R. (2007). An overview of the Tesseract OCR engine. *Ninth International Conference on Document Analysis and Recognition. ICDAR 2007.* Vol. 2. IEEE.

TANNER, S., MUÑOZ, T., & ROS, P. H. (2009). Measuring mass text digitization quality and usefulness. *D-Lib Magazine*, *15*(7/8), 1082-9873.

 © ATALA

Searching LTs for minority languages

Claudia Soria[1] Joseph Mariani[2]
(1) CNR-ILC, via Moruzzi 1, 56124 Pisa (Italy)
(2) LIMSI-CNRS & IMMI, BP133, 91403 Orsay cedex (France)
claudia.soria@ilc.cnr.it, Joseph.Mariani@limsi.fr

RÉSUMÉ

A la recherche de Technologies de la Langue pour les langues minoritaires.

Les Technologies de la Langue (TL) sont un instrument nécessaire pour toutes les langues, en particulier celles qui aspirent à conquérir un espace dans les dispositifs numériques. Les langues qui ne sont pas équipées de technologies sont sérieusement menacées d'extinction numérique dans le long terme. Le projet META-NET a évalué l'état actuel des TL pour 30 langues européennes. Mais qu'en est-il des langues régionales et minoritaires ? Très peu d'informations sont disponibles pour elles. Dans cet article, nous présentons d'abord les données disponibles dans la LRE Map. Nous plaidons ensuite en faveur d'une campagne donnant une image complète des TL existant pour les langues minoritaires et régionales d'Europe. Cela aidera les décideurs, les chercheurs et les développeurs à planifier une feuille de route pour doter toutes les langues des instruments nécessaires pour fonctionner comme des langues correctement équipées dans l'ère numérique. Un recensement des TL disponibles et nécessaires est ainsi proposé.

ABSTRACT

LTs are a necessary instrument for all languages, especially for those aiming at conquering a space over digital devices. Languages that are not equipped with LT seriously face digital extinction in the long run. The META-NET project has assessed the current status of LTs for 30 European languages. But what about regional and minority languages? Very little information is available for them. In this paper, we first present the data available using the LRE Map. We then argue for a campaign aimed at providing a complete picture of LTs for minority and regional languages of Europe. This will help policy makers, researchers and developers to plan a roadmap for doting all languages with the necessary instruments to act as fully equipped languages in the digital era. A census of available and needed LT is thus proposed.

MOTS-CLÉS : Langues régionales et minoritaires, ressources langagières, technologies de la langue.
KEYWORDS: Regional and minority languages, language resources, language technologies.

1 Introduction

1.1 Why do we need LTs for minority languages

Small languages are now using social media, YouTube, text messaging and various technologies to expand their voice and expand their presence. There are several examples of how new media and digital technologies can help the salvation of moribund or endangered languages : North American tribes use social media to re-engage their young, and there is an iPhone app to teach new students the pronunciation of Tuvan words (an

 © ATALA

indigenous tongue spoken by nomadic peoples in Siberia and Mongolia). An app for Tusaalanga Inuktitut is being developed as a resource for learning several Inuktitut dialects. More and more linguistic groups turn to the web as a powerful instrument for preservation and revitalization of their languages. Thousands of people blog and twitter in their mother tongue and use Facebook and Twitter in their languages. The web has become a fantastic opportunity for minority languages, as it offers publishing opportunities at virtually no cost, and under no control. The web is giving back to minority languages the space that was denied to them by other media.

Although not all languages can survive, and many inevitably will be lost as remaining speakers die off, the new digital tools do offer a way back from the brink for a lot of languages that seemed doomed just a few years ago. Small languages can and must profit of the incredible opportunity offered by current technologies.

Facebook, for example, is a wonderful venue for endangered and minority languages communities, especially since it makes those languages attractive in the eyes of the young generation. Social media have become very powerful for language revitalization. From Facebook to voice recognition to interactive learning tools, both small and large communities, private and public, can subvert the tools of daily modern life to keep a language alive.

The role of Information and Communication Technologies (ICTs), thus, is of absolute importance. It will show young people that their community language is up and running, and that it can express all kinds of concepts. Thus, ICTs are powerful mechanisms for building up pride in the language. Internet and, in general, ICTs will help spreading awareness about the problem of language endangerment and preservation, in a capillary way. ICT is not to be seen as a threat, but rather, an opportunity for a small language to extend its voice and to reach a global audience.

At the same time, however, ICTs constitute a big challenge for minority languages, as they require fast development of high quality Language Technologies to keep up the pace of technological development. In other words, ICTs will help minority languages in their gaining a space over the digital place as long as good and effective language technologies are developed and integrated into ICTs: if a language is not adequately supported by language technologies, its use over the Internet and through digital devices becomes cumbersome, communication is difficult, and usability dramatically affected.

This risk affects wider spoken languages as well. Suppose that machine translation becomes more imprecise instead of improving, or advanced voice interfaces are lacking, or typing on a smartphone keyboard become painfully slow because there is no support for a given alphabet: these scenarios can become true if Language Technologies are not adequately developed at a fast pace and for all languages aiming to get hold of the digital space.

The scenario above is what the META-NET Language White Papers[1] (Rehm and Uszkoreit, 2011) refer to with the concept of "digital extinction": when a language is used less and less over the Internet because of lack of LT support, then the range of contexts where it is used dramatically collapses and gradually brings the language to disappear from the digital space.

[1] http://www.meta-net.eu/whitepapers/overview

© ATALA

Language Technologies are today recognized as an enabling technology, allowing languages to be used by humans and machines alike over digital devices. According to the META-NET Language White Papers, all European languages are at risk, with the exception of English. The survey focused on 30 European languages and performed an accurate analysis of the Language Technologies available for them, mostly using the results of the LRE Map. In this paper, we illustrate this tool, its features and potential and argue for its application to regional and minority languages.

1.2 Why do we need to know about LTs

Not only are LTs essential, but knowledge about the missing ones is equally crucial. Knowledge about existing Language Resources (LRs) and Language Technologies (LTs) is essential to the overall advancement of research in the field: it is important to be able to locate and retrieve the right resources for the right applications, and to exploit existing ones before building new ones from scratch. This is particularly true for regional and minority languages, given the limited funding usually available, and the often fragmented framework in which development of LTs takes place. Having a clear picture of which resources are available for which languages and for which use is important in order to identify existing gaps for certain languages at a given time and estimate the amount of investment needed to fill them in. Knowledge about the current use of resources is equally important. Knowing which resources are most used for the various applications will help to better understand the reason behind their success (their intrinsic quality, their wide availability, their licensing model, etc.).

Unfortunately, clear and easy-to-reach information of this type about resources and related technologies is still lacking. Several institutions worldwide maintain catalogues of language resources (ELRA[2], LDC[3], National Institute of Information and Communications Technology (NICT) Universal Catalogue[4], ACL Data and Code Repository[5], OLAC[6], LT World[7], etc.). However, it has been estimated that only 10% of existing resources are known, either through distribution catalogues or via direct publicity by providers (web sites and the like).

The majority of language resources is still poorly documented or not documented at all, and use of metadata elements to describe and document resources is still uncommon and often inconsistent. This represents a serious problem that was repeatedly mentioned by e.g. the FLaReNet recommendations (Soria et al. 2012). Single authors can find it difficult to document their own resources, simply because they can have a hard time deciding the relevant set of metadata elements to be used. Moreover, there is no sufficient awareness about the importance of documentation, which is often disregarded as a useless burden.

2 The LRE Map

The LRE Map was an initiative devised to address both the problem of lack of knowledge

[2] http://catalog.elra.info/
[3] http://www.ldc.upenn.edu/Catalog/
[4] http://facet.shachi.org/?ln=en
[5] http://www.aclweb.org
[6] http://www.language-archives.org/
[7] http://www.lt-world.org/

© ATALA

about existing resources and the need for providing a simple and easy way to encourage documentation of resources. It was jointly launched by FLaReNet (the Fostering Language Resources Network) and ELRA (European Language Resources Association) in May 2010 with the purpose of developing an entirely new instrument for capturing the community knowledge about language resources, collecting descriptions both for tools and existing/ new resources applied in NLP research.

The first Map was initially created in conjunction with the LREC 2010 Conference (see Calzolari et al., 2010), as a campaign for gathering information about the language resources and technologies underlying the scientific works presented during the conference. Authors who submitted a paper were requested to provide information about the language resources and tools either developed or used; the initiative was successful, with close to 2000 resources descriptions. The required information was pretty simple and related to basic information about the type of the resource, the language and modality represented, the intended or real application purposes, the degree of availability for further use, the maturity status, the size, type of license and availability of documentation.

Since the initial LREC Map in 2010, thanks to a massive input from the community, we harvested more data coming from 9 upcoming conferences on Speech and Natural Language Processing (COLING 2010, EMNLP 2010, ACL-HLT 2011, RANLP 2011, Interspeech 2011, Oriental-Cocosda 2011, IJCNLP 2011, LTC 2011 and LREC 2012), which constitute nowadays the LRE Map. The online acquisition facility was improved based on the experience gained in the LREC 2010 Map. More LR Types were suggested, auto-completion was introduced for entering the Type and Name of the LRs, Languages were normalized and the "Speech/Written" Modality was added in the new version, which was used by LREC 2012 authors. However, this data also had to go to a lengthy cleaning operation even if the auto-completion process has improved the homogeneity of the information. This resulted in the LRE Map, which contains 4295 entries, almost 3 times more than in the LREC 2010 Map. Its main goal remains to gather information collected bottom up and to exploit the community knowledge helping the discovery and documentation of resources, essentially through a web interface[8] that enables searches with multiple criteria.

2.1 The LRE Metadata

Each resource is described according to 12 main metadata fields, which provide the minimum set of relevant and useful information about new and existing resources and their use. A set of 9 first metadata has been revised after the abstract submission phase in order to slightly increase the descriptive parameters requested upon submission of the final papers, while an additional set of 3 were added in the final submission phase.

The first set of metadata represents quite general information available in most LR catalogues (such as LDC and ELRA) and surveys (e.g. ENABLER). Each of these basic fields has a list of suggested values, which has been deliberately kept short by using only most frequent and common values. However, the possibility has been left open for the user to select an "Other" field and to specify a more appropriate term in case he/she did not feel any of the suggested values were satisfying. This set contains: a) Resource Type b)

[8] http://www.resourcebook.eu

© ATALA

Resource Name c) Resource Production Status d) Use of the Resource e) Language(s) f) Modality g) Resource Availability h) Resource URL (if available) i) Resource Description. Three descriptors ("Resource Size", "Resource License" and "Resource Documentation") were added in the final submission phase to allow for extraction of additional information.

3 What the LRE Map tells us about European Languages

The LRE Map represents an observatory wherefrom the landscape of language resources and technologies can be observed and analyzed. Although it cannot claim for absolute coverage, it can nevertheless claim, by its size, to be considered as an accurate representation of the current situation.

In this paper we will mostly present an analysis of the European languages (both official and regional/minority ones) represented in the Map and how this data can be interpreted, in particular from the point of view of designing a roadmap for further development of missing resources and of planning language-sensitive strategies for equally (or more evenly) equipping languages with the fundamental tools to function in the digital era.

3.1 Languages and resources

Taking into account European languages only (taken as languages spoken over the European continent, not only languages of the EU), the map offers information for 74 different languages (including, as said before, regional and minority languages and sign languages) and 3187 language resources. The data present a skewed situation with English dominating (1251 resources), second place is for French with 265 resources.

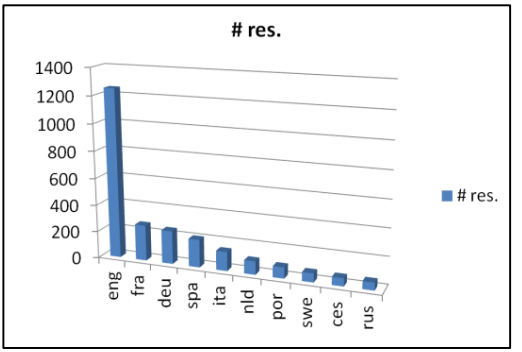


FIGURE 1 – Top ten EU languages

Figures 1-5 illustrate the 10 top languages (Fig. 1), the distribution of resources across EU official languages (Fig. 2-3), and across other languages of the European continent (Fig. 4). The Map also contains information for a total of 148 resources for European regional and minority languages, here presented in a reverse order in terms of number of resources: Basque (45), Catalan (43), Galician (12), Faroese (5), Welsh (4), Aragonese, Asturian, Breton, Frisian, Romansh, Swiss German, and Venetan (3), Luxembourgish, Norwegian Bokmal, Norwegian Nynorsk (2), Corsican, Drents, Limburgan, Lombard, Low German, Lule Saami, North Saami, Occitan, Scottish Gaelic, Scots, Sicilian and South Saami (1).

© ATALA

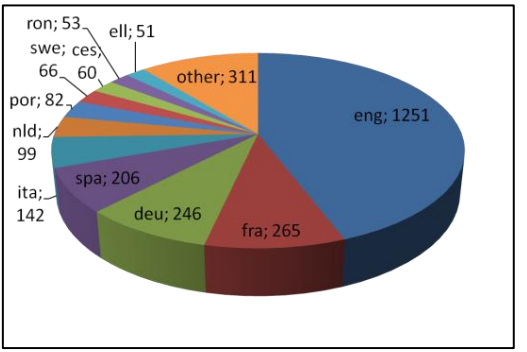


FIGURE 2 – EU languages

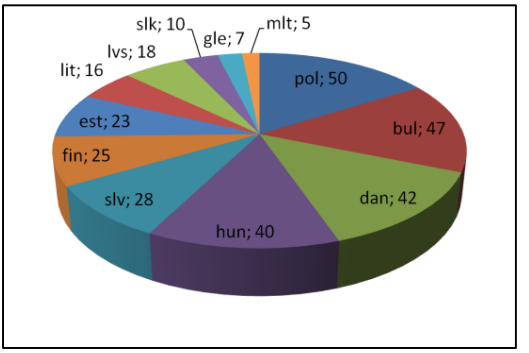


FIGURE 3 – Other EU languages

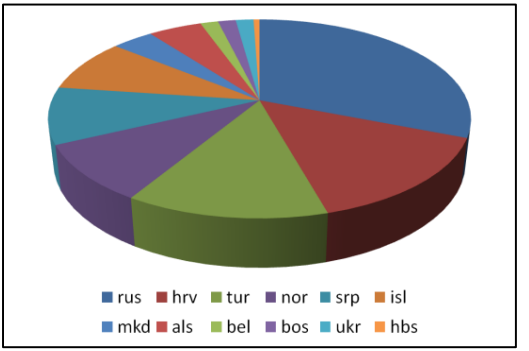


FIGURE 4 – Official non-EU languages

© ATALA

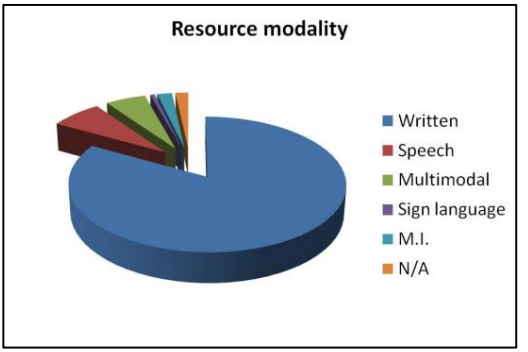


FIGURE 5 – Distribution of EU resources across modalities

3.2 Narrowing the focus: a close-up on resource modality and resource type

We now turn to an illustration to the kind of information that can be derived from other LRE Map metadata, e.g. the Resource Modality and Resource Type category. For the first parameter, the suggested options are « Written », « Speech », « Multimodal/Multimedia », « Sign Language », «Written/Speech » or « Not applicable » (the choice for those types of resources for which modality is not a relevant descriptive parameter, such as tools or ontologies, for instance). Figure 6 illustrates the results for European languages, showing how the vast majority of resources available for European languages belongs to the written modality. If we further analyze the modality parameter distinguishing among EU, non EU and regional/minority languages, we can observe how the same pattern holds true, with written resources being the first and foremost type of resources developed.

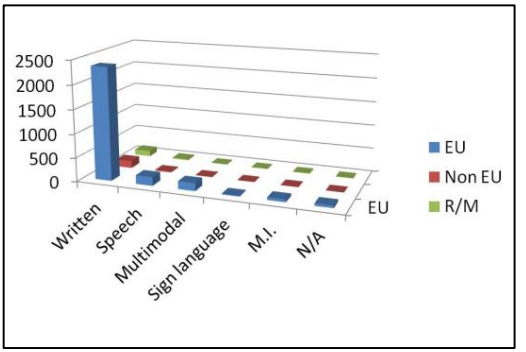


FIGURE 6 – EU, non-EU and European regional/minority language resources across modalities

The "Resource Type" descriptor is aimed at capturing the general type of the resource used or created being described. A wide notion of resource is adopted by the LRE Map that encompasses not only the usual types of resources, such as corpora and lexicons, but also tools, metadata, guidelines and standards, evaluation data, tools and methodologies. On the basis of the data available, it is clear that the majority of resources being developed for

© ATALA

European languages across the various modalities belongs to the macro-category « Data », as illustrated by the table below.

	Written	Spoken	Multimodal
Data	72%	80%	77%
Tools	17%	13%	20%
Evaluation	8%	7%	2%
Guidelines	3%	0%	1%

4 Using the Map

Very detailed analyses can be performed by crossing different parameters. For instance, it is possible to derive the particular types of resources available for each language, or the amount of newly developed resources vs. already existing ones, or else the number of resources that are freely available in contrast with those available for a fee or not available at all. The interested reader is referred to (Calzolari et al. 2010) for an illustration of the various uses the Map lends itself for. Here we are particularly interested in showing how the data gathered with the LRE Map can be used to derive a picture of the particular types of resources available and missing for the various languages, a concept referred to as « Language Matrices ».

4.1 The Language Matrices

The objective of the Language Matrices is to provide a clear picture on what exist in terms of Language Resources, in the broad sense (Data, Tools, Evaluation and Meta-resources (standards, metadata, guidelines)) for the various languages, and to stress the languages that are missing such Language Resources. The goal is then to ensure the production of the corresponding resources to fill the gaps for those languages. In order to do so, we produced Languages Matrices that provide an easy way to get that picture and to have access to the details of the corresponding resources.

A first set of Languages Matrices has been produced in January 2011. Those matrices were built from the LREC Map that has been produced from the information provided by the authors of the papers submitted at the LREC'2010 conference (May 2010), which comprises close to 2,000 entries.

In order to produce the Language Matrices, we had to clean up the data, as the information was not always provided in the proper format, despite the suggested terms, and as the declared LR should only be counted once (see Calzolari et al., 2012). This very tedious process reduced the number of entries to about 1,500.

We wrote a piece of software in order to provide the number of resources that exist for each language, with direct access to all the corresponding information (Mariani et al., 2012). In this first analysis, we considered the 23 official languages of the EU, together with

 © ATALA

a category on "Regional European languages" and one on "Non-EU European languages", as well as "Multilingual", "Language Independent" and "Not Applicable" categories.

We produced 8 Language Matrices on Multimodal/Multimedia Data and Tools, Written Language Data and Tools, Spoken Language Data and Tools, Evaluation and Meta-resources. Several Types of resources are listed for each matrix, either corresponding to the 24 Types that were suggested in the questionnaire or to the author's own wording. This resulted in a total of 160 Language Resource Types, with a variable number for each matrix (from 5 Types for Evaluation to 78 Types for Written Language Tools). Those matrices showed that the English language is by far the most resourced language, followed by French and German, Spanish, Italian and Dutch. Some languages were clearly under-resourced, such as Irish Gaelic, Slovak or Maltese. Given the large number of Types expressed by the authors, some Types may exist only for one language, and the matrices therefore show a large number of zeroes for all the other languages. We however preferred to keep that information as such rather than merging them into an "Other Type" category, as those singletons may be weak signals announcing a new research trend.

We produced the same 8 Language Matrices on the basis of the 2012 LRE Map which includes 216 languages. Also a Language Matrix has been built for Sign Languages, covering 21 Sign Languages. Figures 7-9 present simplified Language Matrices for European languages (respectively EU official, non-EU official and regional or minority languages).

		European Union																						
Resource Modality	Resource Category	Bulgarian	Czech	Danish	Dutch	English	Estonian	Finnish	French	German	Greek	Hungarian	Irish	Italian	Latvian	Lithuanian	Maltese	Polish	Portuguese	Romanian	Slovak	Slovene	Spanish	Swedish
Written	Data	40	35	27	53	738	14	16	144	133	29	28	4	92	15	12	4	29	53	36	9	20	116	42
	Evaluation	2	5	3	6	117		1	10	14	1	1	1	5	3	3			5	3			14	3
	Guidelines	2	4	1	6	32	1		6	6	1	2	2	4				3	1	1		1	7	3
	Tools	3	6	2	12	169	4	2	44	37	12	4		23			1	6	12	10	1	5	29	6
Speech	Data		3	4	6	57		1	27	19	2	1		3				2	4	1		1	14	3
	Evaluation					5			7	1												1		
	Tools		1	1	1	7			1	4		1							1	1			1	2
Multimodal	Data		1	3	6	56	3	3	9	17	4	2		4				3	2				11	4
	Evaluation					3																	1	
	Guidelines					1																		
	Tools		1		1	10		1	3	4	1	1		2		1		1	2	1			3	
Sign language	Data		1		1	4		1	3	1				1									1	1
	Guidelines					1																		
	Tools					1			1															
Written/Speech	Data		2	1	3	19	1		5	4				3				4	1				3	2
	Tools				1	5			1	1													1	
Not applicable	Data				3	9			1	1				2					1				2	
	Guidelines								1	1				1										
	Tools					17			2	3	1			2				2					3	
Total		47	60	42	99	1251	23	25	265	246	51	40	7	142	18	16	5	50	82	53	10	28	206	66

FIGURE 7 – Simplified Language Matrix for the 23 EU official languages

© ATALA

		Europe non EU											
Resource Modality	**Resource Category**	Albanian	Belarusan	Bosnian	Croatian	Icelandic	Macedonian	Norwegian	Russian	Serbian	Serbo-Croatian	Turkish	Ukrainian
Written	Data	8	3	2	21	7	5	9	36	15		13	2
	Evaluation	1			2		1	1	1	2		2	
	Guidelines				1		1		2			1	
	Tools				2	6		2	5			2	
Speech	Data				1	1		4	3			3	1
	Evaluation								1				
	Tools					1			2				
Multimodal	Data							1	3			1	
	Tools											1	
Sign language									1				
Written/Speech	Data			1		1			1			1	
	Tools								1				
Not applicable	Data								1				
Total		**9**	**3**	**3**	**27**	**16**	**7**	**17**	**57**	**17**	**1**	**24**	**3**

FIGURE 8 – Simplified Language Matrix for non-EU languages

		Regional and minority																										
Resource Modality	**Resource Category**	Aragonese	Asturian	Basque	Breton	Catalan	Corsican	Drents	Faroese	Frisian	Galician	Limburgan	Lombard	Low German	Lule Saami	Luxembourgish	North Saami	Norwegian Bokmal	Norwegian Nynorks	Occitan	Romansh	Scottish Gaelic	Scots	Sicilian	South Saami	Swiss German	Venetan	Welsh
Written	Data	2	1	18	3	23	1	1	2	2	7	1	1	1		1		1	2	1	3	1	1	1		1	2	2
	Evaluation			1																								
	Guidelines			2		1																						
	Tools	1	2	12		11			1		2				1		1								1		1	2
Speech	Data			6		5			1		3					1										2		
	Evaluation					1																						
	Tools			3					1									1										
Multimodal	Data					1																						
	Tools			1																								
Sign language	Data					1																						
Written/Speech	Data			1						1																		
Not applicable	Data			1																								
Total		**3**	**3**	**45**	**3**	**43**	**1**	**1**	**5**	**3**	**12**	**1**	**1**	**1**	**1**	**2**	**1**	**2**	**2**	**1**	**3**	**1**	**1**	**1**	**1**	**3**	**3**	**4**

FIGURE 9 –Simplified Language Matrix for European regional and minority languages

The analysis of Language Matrices confirms the supremacy of the (American) English language for all categories of Language Resources. On the 5218 entries, it contains 1364 LR related to English (close to 26%, compared to 29% in 2011), followed by French (286), German (265), Spanish (222), Italian (158) and Dutch (110), while still very few exist for Slovak (9), Irish Gaelic (5) or Maltese (4). However, it appears that the effort is increasing for other major languages (most of them doubled the number of resources), while several previously considered 'minor' languages or regional languages have benefited from a strong

© ATALA

support to recover and some of them (such as Bulgarian, Estonian, Polish or Slovene) more than tripled the number of resources. We note a big increase for the Estonian language (from 7 to 23), for Regional languages (67 to 103) and for non-EU European languages (63 to 293).

Year	Total	English	French	German	Spanish	Italian	Dutch	Estonian	Irish	Slovak	Maltese	EU Regional	Other Europe
2011	1889	559	143	132	111	90	54	7	3	3	2	67	63
2012	5218	1364	286	265	222	158	110	23	5	9	4	103	293

FIGURE 10 - Some examples of evolution from 2011 to 2012.

The Language Matrices derived by the LRE Map also represent an important instrument for assessing the particular types of resources that are missing for any given language, thus usefully complementing, in a data-driven approach, the well known BLaRK (Basic Language Resource Kit) notion (Krauwer, 2003).

For instance, the following figure illustrates the gaps, in terms of missing language resources, for European regional and minority languages on the basis of the available information.

Resource Category	Resource Type	Aragonese	Asturian	Basque	Breton	Catalan	Corsican	Drents	Faroese	Frisian	Galician	Limburgan	Lombard	Low German	Lule Saami	Luxembourgish	North Saami	Norwegian Bokmal	Norwegian Nynorks	Occitan	Romansh	Scottish Gaelic	Scots	Sicilian	South Saami	Swiss German	Venetian	Welsh
Data	Corpus	2	1	10	2	12	1		1	2	4	1	1	1		1		1	2	1	2	1	1	1			2	1
	Lexicon			6		8		1			3										1					1		1
	Ontology																											
	Grammar/Language model					1																						
	Terminology			2		2																						
Tools	Annotation Tool			1		2																						
	Tokenizer		1			3					1																	1
	Tagger/Parser		1	4		3			1		1				1		1								1			1
	Named Entity Recognizer			1																								
	Word Sense Disambiguator			1																								
	Language Identifier																											
	Transcriber																											
	Machine Translation to	1		1																								
	Other			4		3																					1	

FIGURE 11 – Language Matrix for European regional and minority languages

(written resources)

In a similar vein, for instance, the Language Matrices already started being used for identifying the gaps and establishing the Language Tables and Language Status estimate in

© ATALA

the META-NET Language White Papers (Rehm and Uszkoreit, 2012).

In addition to a language and resource coverage analysis, we also introduced a measure of the popularity of the Language Resources, named the Language Resource Impact Factor (LRIF), obtained by counting the number of times a resource is cited. 4 LRIF Matrices have been produced based on the papers accepted at the previously mentioned conferences, which provide a glimpse of the most popular resources in each category (Data, Tools, Evaluation and Meta-Resources).

5 Conclusion

The LRE Map holds an unprecedented potential for possible applications and uses. It is an instrument for enhancing availability of information about resources, either new or already existing ones. It is a measuring tool for monitoring various dimensions of resources across places and times, thus helping highlighting evolutionary trends in language resource use and related language technology development by cataloguing not only language resources in a narrow sense (i.e. language data), but also tools, standards, and annotation guidelines.

The potential of the LRE Map for becoming a powerful aggregator of information related to language resources was immediately clear, as was the possibility of deriving and discovering new combinations of information in entirely new ways. The database underlying the LRE Map can yield interesting matrices of the language resources available for the various languages, modalities, or applications. Such matrices have been already used, for example, in META-NET to provide a picture of the situation of resources availability for the various European languages.

In the near future the LRE Map will continue collecting bottom-up input about resources from authors of other conferences. Providing information about resources could become part of the standard submission process. This will help broadening the notion of "language resources" and thus attract to the field neighboring disciplines that so far have been only marginally involved by the standard notion of language resources. Moreover, it will be extended to authors submitting papers to the "Language Resources and Evaluation" journal.

The LRE Map will have an impact in reinforcing and facilitating the use of standards in the community by allowing registration of resources together with submission of papers for a conference, making most used/most adopted standards emerge. Finally, the LRE Map wants to promote active and personal engagement in documenting resources, encouraging a change in culture.

We therefore strongly advocate the use of the LRE Map by organizations, individual researchers, industry, etc. etc. to document resources for regional and minority languages in order to derive an accurate picture of available technologies and an analysis of the current needs and gaps to be addressed.

References

CALZOLARI N., SORIA C., DEL GRATTA R., GOGGI S., QUOCHI V., RUSSO I., CHOUKRI K., MARIANI J. and PIPERIDIS S. (2010). The LREC Map of Language Resources and Technologies. In *Proceedings of LREC 2010*. La Valletta, Malta, 17-23 May 2010, pp. 949-956.

 © ATALA

CALZOLARI N., DEL GRATTA R., FRANCOPOULO G., MARIANI J., RUBINO F., RUSSO I., and SORIA C. (2012). The LRE Map. Harmonising Community Description of Resources. In *Proceedings of LREC 2012*. Istanbul, Turkey 23-25 May 2012, pp. 1084-1089.

KRAUWER S. (2003). The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In *Proceedings of the 2003 International Workshop Speech and Computer (SPECOM 2003)*, pp. 8-15.

MARIANI J. and FRANCOPOULO G. (2012). The Language Matrices and the Language Resource Impact Factor, Parole Workshop, Lisbon, 18-19 October 2012

REHM, G. and USZKOREIT H., editors (2012). *META-NET White Paper Series*. Springer, ISBN 978-3-642-30775-1.

SORIA C., BEL N., CHOUKRI K., MARIANI J., MONACHINI M., ODIJK J., PIPERIDIS S., QUOCHI V., and CALZOLARI N. (2012). The FLaReNet Strategic Language Resource Agenda. In *Proceedings of LREC'12*. Istanbul, Turkey, 23-25 May 2012, pp. 1379-1386.

© ATALA

Liste des auteurs

A
Adam, Clémentine 140

B
Baroni, Marco 99
Bernhard, Delphine 209
Bertels, Ann 126
Bittar, André 48
Blache, Philippe 108
Bost, Xavier 37
Brixtel, Romain 29
Brunetti, Ilaria 37

C
Cabrera-Diego, Luis Adrián 37
Charton, Eric 75
Clergerie, Éric Villemonte de La 167
Collin, Olivier 61
Constant, Matthieu 95
Cossu, Jean-Valère 37

D
Dini, Luca 48
Dufour, Richard 37

E
El-Bèze, Marc 37

F
Fabre, Cécile 140
Farahmand, Meghdad 92

G
Gagnon, Michel 75
Grabar, Natalia 15
Grouin, Cyril 1
Guerraz, Aleksandra 61

H
Hamon, Thierry 15
Heinecke, Johannes 181
Heylen, Kris 154
Hiou, Yannick 61

J
Jean-Louis, Ludovic 75
Joubert, Alain 103

L
Lafourcade, Mathieu 103
Lecluze, Charlotte 29
Lejeune, Gaël 29
Ligozat, Anne-Laure 209
Linhares, Andréa 37

M
Mariani, Joseph 235
Meurs, Marie-Jean 75
Morardo, Mikaël 167
Morchid, Mohamed 37
Morlane-Hondère, François 112

N
Nerima, Luka 92

P
Paroubek, Patrick 1
Périnet, Amandine 15

R
Rauzy, Stéphane 108
Ruhlmann, Mathieu 48

S
Sagot, Benoît 195
Sandillon-Rezer, Noémie-Fleur 88
Scherrer, Yves 195
Soria, Claudia 235
Speelman, Dirk 126, 154

T
Tanguy, Ludovic 140
Tellier, Isabelle 95
Torres-Moreno, Juan-Manuel 37

U
Urieli, Assaf 221

V
Vecchi, Eva Maria 99
Vergez-Couret, Marianne 221
Voisine, Nicolas 61

W
Wehrli, Eric 92
Wielfaert, Thomas 154

Z
Zamparelli, Roberto 99
Zarrouk, Manel 103
Zweigenbaum, Pierre 1

Liste des Mots clés

A
algorithmique du texte 29
allemand 209
alsacien 209
analyse de cooccurrences 126
analyse distributionnelle 112
analyse syntaxique 92
anaylse de recettes 48
annotation sémantique 15
appariement 29
apprentissage automatique 221

B
bigrammes 92

C
campagne d'évaluation 1, 15
champs markoviens aléatoires 95
classification 1, 29, 75
classification automatique 48
classification de textes 37
classification supervisée 61
clustering hiérarchique 88
co-clustering 61
compositionnalité 99
cooccurrents de deuxième ordre et de troisième ordre 126
corpus 112
création de ressources sémantiques . . . 15

D
deft2013 75
dépendances syntaxiques 167
dialecte 209
dialectes 221
distance sémantique 61
domaine de spécialité 37

E
encoding 181
enrichissement 103
espaces vectoriels sémantiques 154
étiquetage morphosyntaxique . . 195, 209
extraction d'information . . . 29, 37, 48, 75
extraction terminologique 167

F
filtrage statistique 103

G
grammaire ab 88
grammaires catégorielles 88
grammaires de propriétés 108

H
hypothèse distributionnelle 167

I
inférence de règles 103
inférence grammaticale 88
interface web collaborative 167

K
khiops 61

L
langues caucasiennes 181
langues non-dotées 195
langues proches 195
langues régionales et minoritaires . . . 235
lexicographie 154
lexique 221

M
mesures 1
mesures d'association 126
méthodes hybrides 48
méthodes statistiques 37
métriques de distance 126

N
néerlandais 154
niveau d'occurence 154

O
occitan 221
océrisation 221
ocr 221
ordre des mots 99

P
positionnement multidimensionnel . 126, 154

R
recette de cuisine 75
recettes 1
recettes de cuisine 15
recherche d'information 140
reformulation de requêtes 140
réseau lexical 167
réseau lexico-sémantique 103
ressources 181
ressources langagières 235
ressources lexicales et grammaticales . 95

S
satisfaction de contraintes 108
segmentation lexicale 95
segmentation syntaxique 95
sémantique distribuionnelle 99
sémantique distributionnelle . . . 140, 154
sémantique lexicale 112, 154
synonymie 112
système d'apprentissage automatique 15
systèmes à base de règles 15
systèmes hybrides 92

T
taln 181
tchétchène 181
technologies de la langue 235
tilt 61
traitement automatique des langues . . 15
traits symboliques 103
treebank 108

V
visualisation 154

W
wikipédia 61

Y
yiddish 221