



6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition) (JEP-TALN-RÉCITAL) ¹

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition).

Volume 4 : Démonstrations et résumés d'articles internationaux

Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni, Sylvain Pogodalla, Stéphane Schneider (Éds.)

Nancy, France, 08-19 juin 2020

1. <https://jep-taln2020.loria.fr/>

Crédits : L'image utilisée en bannière est une photographie du vitrail « Roses et Mouettes », visible dans la maison Bergeret à Nancy. La [photographie](#) a été prise par Alexandre Prevot, diffusée sur flickr sous la licence [CC-BY-SA 2.0](#).

Le logo de la conférence a été créé par Annabelle Arena.

©2020 ATALA et AFCP

Avec le soutien de



Message des présidents de l’AFCP et de l’ATALA

En ce printemps 2020, et les circonstances exceptionnelles qui l’accompagnent, c’est avec une émotion toute particulière que nous vous convions à la 6e édition conjointe des Journées d’Études sur la Parole (JEP), de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) et des Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL). Après une première édition commune en 2002 (à Nancy, déjà !), et une expérience renouvelée avec succès en 2004, c’est désormais tous les quatre ans (Avignon 2008, Grenoble 2012) que se répète cet événement commun, attendu de pied ferme par les membres des deux communautés scientifiques voisines.

Cette édition 2020 est exceptionnelle, puisque dans le cadre des mesures sanitaires liées à la pandémie mondiale de COVID-19 (confinement strict, puis déconfinement progressif), la conférence ne peut avoir lieu à Nancy comme initialement prévu, mais se déroule à distance, sous forme virtuelle, soutenue par les technologies de l’information et de la communication. Nous remercions ici chaleureusement les organisateurs, Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni et Sylvain Pogodalla, qui ont dû faire preuve de souplesse, d’inventivité, de détermination, de puissance de travail, et de tant d’autres qualités encore, afin de maintenir la conférence dans ces circonstances, en proposant un format inédit. Grâce aux différentes solutions mises en œuvre dans un délai court, la publication des communications scientifiques est assurée, structurée, et les échanges scientifiques sont favorisés, même à distance.

Bien entendu, nous regrettons tous que cette réunion JEP-TALN-RECITAL ne permette pas, comme ses prédécesseurs, de nouer ou renforcer les liens sociaux entre les différents membres de nos communautés respectives – chercheurs, jeunes et moins jeunes, académiques et industriels, professionnels et étudiants – autour d’une passionnante discussion scientifique ou d’un mémorable événement social. . . Notre conviction est qu’il est indispensable de maintenir à l’avenir de tels lieux d’échanges dans le domaine francophone, afin bien sûr de permettre aux jeunes diplômés de venir présenter leurs travaux et poser leurs questions sans la barrière de la langue, mais aussi de dynamiser nos communautés, de renforcer les échanges et les collaborations, et d’ouvrir la discussion autour des enjeux d’avenir, qui questionnent plus que jamais la place de la science et des scientifiques dans notre société.

Lors de la précédente édition, nous nous interrogeons sur les phénomènes et tendances liés à l’apprentissage profond et sur leurs impacts sur les domaines de la Parole et du TAL. Force est de constater que l’engouement pour ces approches dans nos domaines a permis un retour sur le devant de la scène des domaines liés à l’Intelligence Artificielle, animant parfois un débat tant philosophique que technique sur la place de la machine dans la société, notamment à travers le questionnement sur la vie privée de l’utilisateur. Ces questionnements impactent tant la Parole que le TAL, d’une part sur la place de la gestion des données, d’autre part sur les modèles eux-mêmes. Malgré ces questionnements, nous constatons que les acquis et les expertises perdurent, et les nouvelles approches liées à l’apprentissage profond ont permis un rapprochement des domaines de la Parole et du TAL, sans les dénaturer, à la manière des conférences JEP-TALN-RECITAL qui créent un espace plus grand d’échange et d’enrichissement réciproques.

Nous terminons ces quelques mots d’ouverture en remerciant l’ensemble des personnes qui ont rendu possible cet événement qui restera, nous l’espérons, riche et passionnant, malgré les circonstances. L’ATALA et l’AFCP tiennent tout d’abord à réitérer leurs remerciements aux organisateurs des JEP, de TALN et de RECITAL, qui sont parvenus à maintenir le cap à travers vents et marées. Nos remerciements vont également à l’ensemble des membres des comités de programme, dont le travail et l’implication ont permis de garantir la qualité et la cohérence du programme finalement retenu. Un grand merci aux relecteurs pour le temps et le soin qu’ils ont dédiés à ce travail anonyme et indispensable. Ils se reflètent dans la qualité des soumissions que chacun pourra découvrir sur le site de la conférence.

En conclusion, cette 6e édition conjointe JEP-TALN-RECITAL est exceptionnelle parce qu’elle se tient dans un contexte de crise généralisée — crise sanitaire, économique, voire sociale et politique. Mais nous

formons le vœu qu'elle reste également dans les annales pour la qualité des échanges scientifiques qu'elle aura suscités, et pour le message envoyé à nos communautés scientifiques et à la société dans son ensemble, un message de détermination et de confiance en l'avenir, où la science et les nouvelles technologies restent au service de l'humain.

Véronique Delvaux, présidente de l'Association Francophone de la Communication Parlée
Christophe Servan, président de l'Association pour le Traitement Automatique des Langues

Préface

En 2002, l'[AFCP](#) (Association Francophone pour la Communication Parlée) et l'[ATALA](#) (Association pour le Traitement Automatique des Langues) organisèrent conjointement leurs principales conférences afin de réunir en un seul lieu, à Nancy, les communautés du traitement automatique et de la description des langues écrites, parlées et signées.

En 2020, la sixième conférence commune revient à Nancy, après Fès (2004), Avignon (2008), Grenoble (2012) et Paris (2016). Elle est organisée par le [LORIA](#) (Laboratoire lorrain de recherche en informatique et ses applications, UMR 7503), l'[ATILF](#) (Analyse et traitement informatique de la langue française, UMR 7118) et l'[INIST](#) (Institut de l'information scientifique et technique) et regroupe :

- les 33^{es} Journées d'Études sur la Parole (JEP),
- la 27^e conférence sur le Traitement Automatique des Langues Naturelles (TALN),
- la 22^e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL).

Les circonstances particulières liées à l'épidémie de Covid-19 en France et dans le monde ont conduit à une virtualisation de la conférence. Ainsi, malgré un rassemblement physique qui n'a pu avoir lieu, diffusions, présentations (au gré des auteurs) et discussions des articles acceptés ont lieu sur le site internet de la conférence. Les tutoriels, certains ateliers, et le salon de l'innovation qui accompagnent la conférence ont cependant dû être annulés, mais les ateliers suivants sont maintenus :

- Défi Fouille de Textes (DEFT 2020),
- Éthique et TRaitemeNt Automatique des Langues (ÉTeRNAL).

La conférence accueille également des conférencières et conférenciers invités dont les exposés sont diffusés sur le site : Dirk Hovy (université de Bocconi, Milan, Italie, invité ÉTeRNAL) ainsi que Marie-Jean Meurs (Université du Québec à Montréal, UQAM, Canada) et Hugo Cyr (Faculté de science politique et droit à l'Université du Québec à Montréal, UQAM, Canada). En raison des circonstances particulières, un exposé conjoint de Christine Meunier (Laboratoire Parole et Langage LPL, CNRS, Aix-en-Provence, France) et Christophe Stécoli (police technique et scientifique française) a dû être annulé et reporté à une journée spéciale en septembre 2020.

Ces actes regroupent les articles des conférences JEP (volume 1), TALN (volume 2), RÉCITAL (volume 3), les articles décrivant les démonstrations (volume 4), et les articles des ateliers DEFT (volume 5) et ÉTeRNAL (volume 6). Pour la première fois, un appel spécifique à résumés en français d'articles parus dans une sélection de conférences internationales en 2019 était également proposé (volume 4). Un appel spécifique apprenti·e·s chercheur·euse·s destiné aux étudiants de licence, de master, ou en première année de thèse a également été proposé, pour leur proposer des présentations courtes ou sous forme de poster de leurs projets.

Pour les JEP, 87 articles ont été soumis, parmi lesquels 74 ont été sélectionnés, soit un taux de sélection de 85%.

Pour TALN, 58 articles ont été soumis, parmi lesquels 37 ont été sélectionnés, soit un taux de sélection de 63%, dont 10 comme article longs (17% des soumissions) et 27 comme article courts dont 20 en présentation orale (34% des soumissions) et 7 en présentation poster (12% des soumissions).

Pour RÉCITAL, 22 articles ont été soumis, parmi lesquels 16 ont été sélectionnés, soit un taux de sélection de 73%.

Nous souhaitons vivement remercier toutes les personnes qui ont participé à ce travail de relecture et de sélection :

- l'ensemble des relecteurs (voir page [xi](#)),
- le comité de programme des JEP (voir page [viii](#)),
- le comité de programme de TALN (voir page [ix](#)),
- le comité de programme de RÉCITAL (voir page [x](#)).

Nous souhaitons également remercier nos sociétés savantes : l'AFCP, assurant la continuité des éditions successives des JEP, et l'ATALA, dont le CPerm (comité permanent) assure la continuité des éditions

successives de TALN.

Nous remercions le comité d'organisation et les nombreuses personnes qui ont assuré le soutien administratif et technique pour que cette conférence se déroule dans les meilleures conditions, et en particulier Yannick Parmentier pour son travail pour la diffusion de ces actes sur HAL et les différents sites d'archives ouvertes ([anthologie ACL](#) et [talnarchives.atala.org/](#)).

Nous remercions enfin tous les partenaires institutionnels et industriels qui nous ont fait confiance, en particulier l'université de Lorraine, le CNRS, l'Inria, le LORIA, l'ATILF, l'INIST, le master TAL de l'Institut des Sciences du Digital Management & Cognition (IDMC), le projet OLKI de l'initiative Lorraine Université d'Excellence (LUE), la Région Grand Est, *The Evaluations and Language resources Distribution Agency* (ELDA), le projet ANR PARSEME-FR, la délégation générale à la langue française et aux langues de France (DGLFLF), l'Association des Professionnels des Industries de la Langue (APIL) et les entreprises Synapse, Yseop et Orange.

Bonne conférence à toutes et à tous !

Les présidentes et présidents JEP :	David Langlois et Slim Ouni
TALN :	Chloé Braud et Sylvain Pogodalla
RÉCITAL :	Christophe Benzitoun et Laurine Huber

Comités

Comité de programme des JEP

Martine Adda-Decker (Laboratoire de Phonétique et Phonologie, CNRS)
Jean-Francois Bonastre (LIA, Université d'Avignon)
Fethi Bougares (LIUM, Le Mans Université) Philippe Boula De Mareüil (LIMSI, CNRS)
Hervé Bredin (LIMSI, CNRS)
Olivier Crouzet (LLING, Université de Nantes)
Elisabeth Delais-Roussarie (LLING, Université de Nantes)
Véronique Delvaux (Laboratoire de Phonétique,IRSTL, Université de Mons)
Camille Fauth (LiLPa, Université de Strasbourg)
Emmanuel Ferragne (CLILLAC-ARP, Université de Paris)
Cecile Fougeron (Laboratoire de Phonétique et Phonologie, CNRS)
Corinne Fredouille (LIA, Université d'Avignon)
Alain Ghio (LPL, CNRS)
Camille Guinaudeau (LIMSI, Université Paris Sud)
Anne Guyot Talbot (CLILLAC-ARP, Université de Paris 7)
Bernard Harmegnies (Laboratoire de Phonétique,IRSTL, Université de Mons)
Nathalie Henrich Bernardoni (Gipsa-lab, CNRS)
Bassam Jabaian (LIA, Université d'Avignon)
David Langlois (LORIA, Université de Lorraine)
Yves Laprie (LORIA, CNRS)
Anthony Larcher (LIUM, Université du Maine)
Gwénolé Lecorvé (IRISA, Université de Rennes)
Benjamin Lecouteux (LIG, Université Grenoble Alpes)
Georges Linarès (LIA, Université d'Avignon)
Damien Lolive (IRISA, Université Rennes)
Julie Mauchair (IRIT)
Yohann Meynadier (LPL, Aix-Marseille Université)
Slim Ouni (LORIA, Université de Lorraine)
Thomas Pellegrini (IRIT, Université de Toulouse)
François Portet (LIG, Grenoble INP)
Fabian Santiago (Structures Formelles du Langage, Université de Paris 8)
Christophe Savariaux (Gipsa-lab, CNRS)
Nathalie Vallee (Gipsa-lab, Université Grenoble Alpes)
Ioana Vasilescu (LIMSI, CNRS)

Comités de programme TALN

Maxime Amblard (LORIA, Université de Lorraine)
Chloé Braud (IRIT, CNRS)
Caroline Brun (Naver Labs Europe)
Nathalie Camelin (LIUM, Université du Maine)
Marie Candito (Université Paris 7)
Vincent Claveau (IRISA, CNRS)
Chloé Clavel (Telecom-ParisTech)
Mathieu Constant (ATILF, CNRS, Université de Lorraine)
Pascal Denis (Inria)
Cécile Fabre (Université Toulouse 2)
Thomas François (Université catholique de Louvain)
Núria Gala (LPL, CNRS, Aix-Marseille Université)
Natalia Grabar (STL, CNRS, Université Lille 3)
Anne-Laure Ligozat (LIMSI, CNRS, ENSIE, Université Paris-Saclay)
Emmanuel Morin (LINA, Université de Nantes)
Sylvain Pogodalla (LORIA, Inria)
Solen Quiniou (LINA, Université de Nantes)
Corentin Ribeyre (Etermind)
Tim van de Cruys (IRIT, CNRS)
Pierre Zweigenbaum (LIMSI, CNRS, Université Paris-Saclay)

Comité de programme RÉCITAL

Jean-Yves Antoine (Université François Rabelais de Tours)
Sonia Badene (Linagora, IRIT)
Frédéric Béchet (LIF, Aix Marseille Université)
Christophe Benzitoun (ATILF, Université de Lorraine)
Maria Boritchev (LORIA, Inria)
Léo Bouscarrat (EURA NOVA, Aix-Marseille Université)
Manon Cassier (INALCO, Paris)
Kevin Deturck (Viseo Technologies)
Emmanuelle Esperança-Rodier (GETALP, Université Grenoble Alpes)
Kim Gerdes (sorbonne nouvelle)
Nicolas Hernandez (LINA, UMR 6241, CNRS, Université de Nantes)
Lydia-Mai Ho-Dac (CLLE-ERSS, Université Toulouse Jean Jaurès)
Laurine Huber (LORIA, Université de Lorraine)
Sylvain Kahane (Modyco, Université Paris Ouest Nanterre)
Gwénolé Lecorvé (IRISA, Université de Rennes, CNRS)
Joël Legrand (LORIA, Inria, CNRS)
Anne-Laure Ligozat (LIMSI, CNRS, ENSIE, Université Paris-Saclay)
Pierre Ludmann (LORIA, Université de Lorraine)
Yann Mathet (Université de Caen)
Anne-Lyse Minard (IRISA, CNRS)
Sandrine Ollinger (ATILF, UMR 7118, CNRS)
Yannick Parmentier (LORIA, Université de Lorraine)
Justine Reynaud (LORIA, Université de Lorraine)
Stella Zevio (LIPN, Université de Paris 13)

Relectrices et relecteurs

- Gilles Adda (LIMSI, CNRS) Salah Ait-Mokhtar (Naver Labs Europe)
Charlotte Alazard (Université Toulouse 2 Jean Jaurès)
Alexandre Allauzen (LIMSI-CNRS, Université Paris-Sud)
Pascal Amsili (Université Paris Diderot)
Pierre André Hallé (Laboratoire de Phonétique et Phonologie, CNRS–Université Paris 3)
Régine André-Obrecht (Université Paul Sabatier Toulouse III)
Jean-Yves Antoine (Université François Rabelais de Tours)
Nicolas Audibert (Laboratoire de Phonétique et Phonologie, CNRS–Université Paris 3)
Nelly Barbot (IRISA, Université de Rennes 1)
Claude Barras (LIMSI, CNRS)
Loïc Barrault (University of Sheffield)
Katarina Bartkova (ATILF, Université de Lorraine)
Frédéric Béchet (LIF, Aix Marseille Université)
Nathalie Bedoin (DDL, Université Lyon 2)
Patrice Bellot (LSIS, CNRS, Aix-Marseille Université)
Asma Ben Abacha (National Library of Medicine, National Institutes of Health)
Delphine Bernhard (LiLPa, Université de Strasbourg)
Roxane Bertrand (LPL, CNRS, Aix-Marseille Université)
Laurent Besacier (Laboratoire d’Informatique de Grenoble)
Yves Bestgen (F.R.S-FNRS et Université Catholique de Louvain)
Frédéric Bimbot (IRISA, CNRS)
Caroline Bogliotti (MODYCO, UMR 7114, CNRS, Université Paris Nanterre)
Anne Bonneau (LORIA, CNRS)
Stéphanie Borel (Université de Tours)
Féthi Bougarès (LIUM, Le Mans Université)
Leila Boutora (Laboratoire Parole et Langage, Aix Marseille Université)
Paul Caillon (LORIA, Université de Lorraine)
Mélanie Canault (DDL, Université Lyon 2)
Thierry Charnois (LIPN, CNRS, Université de Paris 13)
Chloé Clavel (Telecom-ParisTech)
Maximin Coavoux (Université Grenoble Alpes, CNRS)
Vincent Colotte (LORIA, Université de Lorraine)
Juan Manuel Coria (LIMSI, Université Paris-Saclay Paris 13)
Benoît Crabbé (Université Paris 7)
Lise Crevier Buchman (Laboratoire de Phonétique et Phonologie, CNRS, Hôpital Foch)
Béatrice Daille (LINA, Université de Nantes)
Géraldine Damnati (Orange Labs)
Dan Dediu (Dynamique du Langage, UMR5596, Université Lumière Lyon 2)
Joseph Di Martino (LORIA, Université de Lorraine)
Gaël Dias (Université Caen Normandie)
Amazouz Djegdjiga (LPP, Université Sorbonne Nouvelle – Paris 3)
Benjamin Elie (IMSIA, ENSTA ParisTech)
Iris Eshkol-Taravella (Université d’Orléans)
Emmanuelle Esperança-Rodier (GETALP, Université Grenoble Alpes)
Yannick Estève (LIA, Université d’Avignon)
Dominique Estival (Western Sydney University)
Olivier Ferret (CEA LIST)
Lionel Fontan (Archean Labs)
Karën Fort (Sorbonne Université)
Claire Gardent (LORIA, CNRS)
Eric Gaussier (LIG, Université Grenoble Alpes)
Cédric Gendrot (LPP, Université Sorbonne Nouvelle – Paris 3)
James German (Laboratoire Parole et Langage, Aix Marseille Université)
Cyril Goutte (National Research Council Canada)
Cyril Grouin (LIMSI, CNRS, Université Paris-Saclay)
Pierre André Hallé (LPP, Université Sorbonne Nouvelle – Paris 3)
Olivier Hamon (Syllabs)
Thierry Hamon (LIMSI, Université Paris-Saclay, CNRS, Université Sorbonne Paris Nord)
Bernard Harmegnies (Institut de Recherche en Sciences et Technologies du Langage, Université de Mons)
Nabil Hathout (CLLE, CNRS)
Amir Hazem (LS2N, Université de Nantes)
Nicolas Hernandez (LS2N, Université de Nantes)
Fabrice Hirsch (Praxiling, Université Paul Valéry Montpellier 3)
Thomas Hueber (GIPSA-lab, CNRS)
Kathy Huet (Institut de Recherche en Sciences et Technologies du Langage, Université de Mons)

Stéphane Huet (LIA, Université d'Avignon)
 Mathilde Hutin (LIMSI, Université Paris-Saclay, CNRS, Université Sorbonne Paris Nord)
 Irina Illina (LORIA, Université de Lorraine)
 Christine Jacquin (LS2N Université de Nantes)
 Adèle Jatteau (STL, UMR 8163, Université de Lille, CNRS)
 Denis Jouvét (LORIA, Inria)
 Sylvain Kahane (Modyco, Université Paris Ouest Nanterre)
 Takeki Kamiyama (LPP, Université Paris 8 Vincennes-Saint-Denis)
 Hannah King (CLILLAC-ARP, Université Paris Diderot)
 Olivier Kraif (Université Grenoble Alpes)
 Matthieu Labeau (Telecom Paris)
 Mathieu Lafourcade (LIRMM, Université de Montpellier)
 Mohamed Lahrouchi (SFL, UMR 7023, CNRS Université Paris 8)
 Muriel Lalain (LPL, CNRS, Aix-Marseille Université)
 Joseph Lark (Dictanova)
 Thomas Lavergne (LIMSI, CNRS, Univ. Paris Sud, Université Paris Saclay)
 Guillaume Le Berre (LORIA, Université de Lorraine)
 Gwénolé Lecorvé (IRISA, Université de Rennes, CNRS)
 Benjamin Lecouteux (Laboratoire Informatique de Grenoble)
 Claire Lemaire (Université Grenoble Alpes)
 Yves Lepage (Waseda University)
 Joseph Le Roux (LIPN, Université de Paris 13)
 Veronika Lux (ATILF, CNRS)
 Paolo Mairano (STL, UMR 8163, Université de Lille)
 Anna Marczyk (LPL, CNRS, Aix-Marseille Université)
 Denis Maurel (Université François Rabelais de Tours)
 Christine Meunier (LPL, CNRS, Aix-Marseille Université)
 Alexis Michaud (LACITO, CNRS)
 Richard Moot (LIRMM, CNRS)
 Véronique Moriceau (LIMSI, CNRS)
 Philippe Muller (IRIT, Université de Toulouse)
 Alexis Nasr (LIF, Université de la Méditerranée)
 Sylvain Navarro (CLLE-ERSS, CNRS)
 Luka Nerima (Université de Genève)
 Aurélie Névél (LIMSI, CNRS, Université Paris-Saclay)
 Jian-Yun Nie (Université de Montréal)
 Damien Nouvel (INaLCO)
 Nicolas Obin (IRCAM)
 Yannick Parmentier (LORIA, Université de Lorraine)
 Sebastian Peña Saldarriaga (Dictanova)
 Marie Philippart de Foy (Université de Mons)
 Myriam Piccaluga (Institut de Recherche en Sciences et Technologies du Langage, Université de Mons)
 Claire Pillot-Loiseau (LPP, UMR 7018, CNRS, Université Sorbonne Nouvelle – Paris 3)
 Serge Pinto (LPL, CNRS, Aix-Marseille Université)
 Agnès Piquard (LORIA, CNRS, Université de Lorraine)
 Thierry Poibeau (LaTTiCe, CNRS)
 Alain Polguère (ATILF Université de Lorraine)
 Laurent Prévot (LPL, CNRS, Aix-Marseille Université)
 Jean-Philippe Prost (LIRMM, Université de Montpellier)
 Christian Raymond (IRISA, INSA de Rennes)
 Christian Retoré (LIRMM, Université de Montpellier)
 Albert Rilliard (LIMSI, CNRS, Université Paris-Saclay)
 Virginie Roland (Institut de Recherche en Sciences et Technologies du Langage, Université de Mons)
 Sophie Rosset (LIMSI, CNRS, Université Paris-Saclay)
 Véronique Sabadell (LPC, Aix Marseille Université)
 Stéphane Schneider (INIST, CNRS)
 Didier Schwab (Université Grenoble Alpes)
 Pascale Sébillot (IRISA, INSA de Rennes)
 Djamé Seddah (Almanach, Université Paris la Sorbonne)
 Gilles Serasset (LIG, Université Grenoble Alpes)
 Romain Serizel (LORIA, Université de Lorraine)
 Kamel Smaïli (LORIA, Université de Lorraine)
 Rudolph Sock (LiLPa, Université de Strasbourg)
 Ludovic Tanguy (CLLE, CNRS)
 Xavier Tannier (LIMICS, Sorbonne Université, INSERM)
 Andon Tchechmedjiev (IMR, Mines Alès)
 Juan-Manuel Torres-Moreno (LIA, Université d'Avignon)
 Nicolas Turenne (LISIS, INRA)
 Béatrice Vaxelaire (LiLPa, Université de Strasbourg)

Anne Vilain (GIPSA-lab, Université de Grenoble Alpes)

Coriandre Vilain (GIPSA-lab, Université de Grenoble Alpes)

Guillaume Wisniewski (LLF, Université de Paris)

Jane Wottawa (LIUM, Le Mans Université)

Yaru Wu (LPP, MoDyCo, Université Paris Nanterre)

Kossi Seto Yibokou (LiLPa, Université de Strasbourg)

François Yvon (LIMSI, CNRS, Université Paris-Sud)

Table des matières

I	Démonstrations	1
	Démo de AMALD-serveur et AMALD-corpus, dédiés à l'analyse morphologique de l'allemand	2
	<i>Christian Boitet, Vincent Berment, Jean-Philippe Guilbaud, Claire Lemaire</i>	
	Générateur de dictionnaires au format Android pour les langues peu dotées	6
	<i>Rémy Bonnet, Guillaume Segerer</i>	
	Désidentification de données texte produites dans un cadre de relation client	10
	<i>Guillaume Dubuisson Duplessis, Elliot Bartholme, Sofiane Kerroua, Mathilde Poulain, Ahès Roulier, Anne-Laure Guénet</i>	
	Analyse sémantique de transcriptions automatiques d'appels téléphoniques en français	14
	<i>Emmanuelle Dusserre, Ruslan Kalitvianski, Mathieu Ruhlmann, Muntsa Padró</i>	
	Muzeeglot : annotation multilingue et multi-sources d'entités musicales à partir de représentations de genres musicaux	18
	<i>Elena V. Epure, Guillaume Salha, Félix Voituret, Marion Baranes, Romain Hennequin</i>	
	FlexEval, création de sites web légers pour des campagnes de tests perceptifs multimédias	22
	<i>Cédric Fayet, Alexis Blond, Grégoire Coulombel, Claude Simon, Damien Lolive, Gwénolé Lecorvé, Jonathan Chevelu, Sébastien Le Maguer</i>	
	Vers une analyse automatique de la perception relative à un lieu	26
	<i>Hélène Flamein, Iris Eshkol-Taravella</i>	
	Un prototype en ligne pour la prédiction du niveau de compétence en anglais des productions écrites	30
	<i>Thomas Gaillat, Nicolas Ballier, Annanda Sousa, Manon Bouyé, Andrew Simpkin, Bernardo Stearns, Manel Zarrouk</i>	
	LEX : un jeu avec finalité d'acquisition de ressources lexicales	34
	<i>Asheesh Gulati</i>	
	SkiF Maker : un processus de classement du meilleur talent par rapport à une offre de mission	38
	<i>Jihen Karoui, Kamilia Hafid, Stéphane Pezeril</i>	
	Iagotchi : vers un agent conversationnel artistique	42
	<i>Frejus Laleye, Gaël de Chalendar, Léopold Frey, Rocio Berenguer</i>	
	Bien choisir son outil d'extraction de contenu à partir du Web	46
	<i>Gaël Lejeune, Adrien Barbaresi</i>	
	Conception d'un système de détection d'intention pour un moteur de recherche sur Internet	50
	<i>Estelle Maudet, Christophe Servan</i>	
	Projet AMIS : résumé et traduction automatique de vidéos	53
	<i>Mohamed Amine Menacer, Dominique Fohr, Denis Jouvét, Karima Abidi, David Langlois, Kamel Smaïli</i>	

La résolution d'anaphores au-delà de la frontière de la phrase	57
<i>Luka Nerima, Eric Wehrli</i>	
Spiderlex et compagnie	60
<i>Sandrine Ollinger, Alain Polguère, Yannick Chudy, Bruno Gaume</i>	
Utiliser les outils CORLI de conversion TEI pour l'analyse de corpus de langage oral	64
<i>Christophe Parisse, Loïc Liégeois</i>	
Vers un corpus optimal pour la fouille de textes : stratégie de constitution de corpus spécialisés à partir d'ISTEX	66
<i>Camille de Salabert, Sabine Barreaux</i>	
II Résumés d'articles acceptés dans des conférences internationales	70
Analyse sémantique robuste par apprentissage antagoniste pour la généralisation de domaine	71
<i>Gabriel Marzinotto, Géraldine Damnati, Frédéric Béchet, Benoît Favre</i>	

Première partie

Démonstrations

Démo de AMALD-serveur et AMALD-corpus, dédiés à l'analyse morphologique de l'allemand

Christian Boitet¹, Vincent Berment^{1,3}, Jean-Philippe Guilbaud¹, Claire Lemaire^{1,2}

(1) LIG-GETALP, UGA, IMAG, 700 av. Centrale, 38058 Grenoble cedex 9, France

(2) LAIRDIL, IUT, Univ. Paul Sabatier, 115 B rte de Narbonne, 31077 Toulouse, France

(3) INALCO, 65 rue des Grands Moulins, 75214 Paris cedex 13, France

{Prénom.Nom}@imag.fr

RÉSUMÉ

Le projet AMALDarium vise à offrir sur la plateforme lingwarium.org (1) un service d'analyse morphologique de l'allemand (AMALD-serveur), à grande couverture et de haute qualité, traitant la flexion, la dérivation et la composition, ainsi que les verbes à particule séparable séparée (ou agglutinée), (2) un corpus de référence de haute qualité donnant tous les résultats possibles de l'analyse morphologique, avant filtrage par une méthode statistique ou syntaxique, et (3) une plateforme (AMALD-éval) permettant d'organiser des évaluations comparatives, dans la perspective d'améliorer les performances d'algorithmes d'apprentissage en morphologie. Nous présentons ici une démonstration en ligne seulement de AMALD-serveur et AMALD-corpus. Le corpus est un sous-ensemble anonymisé et vérifié d'un corpus en allemand formé de textes sur le cancer du sein, contenant de nombreux mots composés techniques.

ABSTRACT

Demonstration of AMALD-serveur and AMALD-corpus, dedicated to the morphological analysis of German

The AMALDarium project aims to offer on the lingwarium.org platform (1) a large-coverage and high-quality morphological analysis service for German (AMALD-server), handling flexion, derivation and composition, as well as verbs with separated (or agglutinated) separable particles, (2) a high-quality reference corpus giving all possible results of the morphological analysis, before filtering by a statistical or syntactic analysis, and (3) a platform (AMALD-eval) to organize comparative evaluations, with a view to improve the performance of morphology learning algorithms. We present an online demonstration of AMALD-server and AMALD-corpus only. The corpus is an anonymized and verified subset of a German corpus of texts on breast cancer, containing many technical compound words. The parser accepts as input a text of any length.

MOTS-CLES : Allemand, analyse morphologique, corpus de référence, services web gratuits.

KEYWORDS: German language, morphological analysis, reference corpus, free web services.

1 Motivations

Dans une publication à TALN-2013 (Guilbaud & al. 2013), nous avons décrit l'AM (analyseur morphologique) AMALD construit par J.P. Guilbaud. Depuis, sa couverture est passée de 103.000 à 209.325 lemmes simples et composés, soit plus de 1M de formes, et le traitement des balises XML (sans liste attributs-valeurs) a été introduit. Il est construit par une méthode « experte » et non pas par une méthode empirique, statistique ou neuronale. En 2013, nous l'avions comparé à d'autres AM en ligne, comme DEMorphy (Altinok 2018), *Disambiguator* (Lezius, Rapp et Wettler, 1998), SMOR et Morphisto (Piskorski *et al.*, 2009). Nous avons montré qu'il était de très loin meilleur, en particulier pour la lemmatisation des formes verbales et des noms composés, et 2

à 3 fois plus couvrant. Par exemple, le corpus de Morphy (annoté par ses résultats), qui en fait n'est pas un corpus mais une liste de mots-formes, montrait 90% de résultats faux au niveau des lemmes sur les formes verbales du corpus associé. D'autre part, AMALD était le seul à traiter convenablement les mots composés connexes, c'est-à-dire à bien les segmenter et à fournir pour chaque morceau un lemme et les informations lexicales associées. AMALD était également le seul AM à regrouper un verbe composé d'un verbe simple et d'une particule séparable, quand la particule n'est pas agglutinée, mais séparée par un nombre quelconque de mots. Nous sommes en train d'actualiser notre étude de l'état de l'art, en y incluant Fips-de du LATL (Scherrer 2008) et phpMorphy. Cela devrait faire l'objet d'une publication ultérieure liée à AMALD-eval.

La construction manuelle d'un AM par un ou des experts est un très gros travail, et on ne peut sans doute pas le poursuivre indéfiniment. Nous avons inclus (comme annoncé en 2013) les ≈210.000 lemmes simples ou composés (agglutinés, comme *Onkogen*, oncogène) trouvés dans le Duden, mais, pour pouvoir monter à 6 ou 7 millions d'entrées (comme dans le système japonais↔anglais ATLAS-2.v13 de Fujitsu), il semble nécessaire de mettre en œuvre des techniques d'apprentissage. Or, pour donner de bons résultats, il faut qu'elles disposent d'un corpus annoté (par les résultats souhaités) de grande taille et de très haute qualité, exactement comme pour la TA ou d'autres applications. Nous avons donc entrepris de construire un tel corpus, avec dans l'idée (1) de partir d'un corpus réel, appartenant à un sous-langage justifiant de nombreuses applications, (2) d'utiliser AMALD pour obtenir une première version des annotations, puis (3) de les évaluer (au niveau des lemmes, des catégories morphosyntaxiques, de la décomposition pour les composés, et des variables d'actualisation comme cas, genre, nombre, personne, temps, mode), et d'améliorer AMALD. Le corpus actuel est un sous-ensemble anonymisé et vérifié de 20 textes tirés d'un corpus formé de textes en allemand sur le cancer du sein, contenant de nombreux mots composés techniques. Nous le présentons plus bas (pas très en détail, à cause de la limite à 4 pages).

Depuis 2013, la couverture de l'AM a été beaucoup augmentée, de sorte qu'il n'y a plus de « mots inconnus » dans ce corpus (à part les balises Xml et les emprunts à l'anglais), et on ne trouve pratiquement pas d'erreur dans les résultats. Nous nous sommes rendu compte qu'AMALD devrait être utilisé par d'autres pour que des erreurs ou incomplétudes puissent nous être signalées, et que nous puissions continuer à corriger ou compléter. Enfin, il est apparu qu'il était très difficile de définir des mesures de qualité pour les AM. C'est dans ce contexte que nous avons défini le projet AMALDarium. Ce projet vise à offrir sur la plate-forme lingwarium.org¹ (1) un service d'analyse morphologique de l'allemand (AMALD-serveur), à grande couverture et de haute qualité, traitant la flexion, la dérivation et la composition, ainsi que les verbes à particule séparable séparée (ou agglutinée), (2) un corpus de référence de haute qualité donnant tous les résultats possibles de l'analyse morphologique, avant filtrage par une méthode statistique ou syntaxique, et (3) une plateforme (AMALD-évaluation) permettant d'organiser des évaluations comparatives, dans la perspective d'améliorer les performances d'algorithmes d'apprentissage en morphologie. Nous présentons ici une démonstration en ligne uniquement de AMALD-serveur et de AMALD-corpus.

2 AMALD-serveur

Il est accessible à <http://51.255.118.18/TestAnaALD/>. On peut entrer un texte complet, codé en UTF-8, en format texte simple ou Xml/TEI simplifié pour l'instant : les balises ouvrantes ne doivent pas contenir de liste d'attributs-valeurs. D'autre part, les balises doivent être séparées du texte (pas '`<h1>Titel`' mais '`<h1> Titel`'). Exemple :

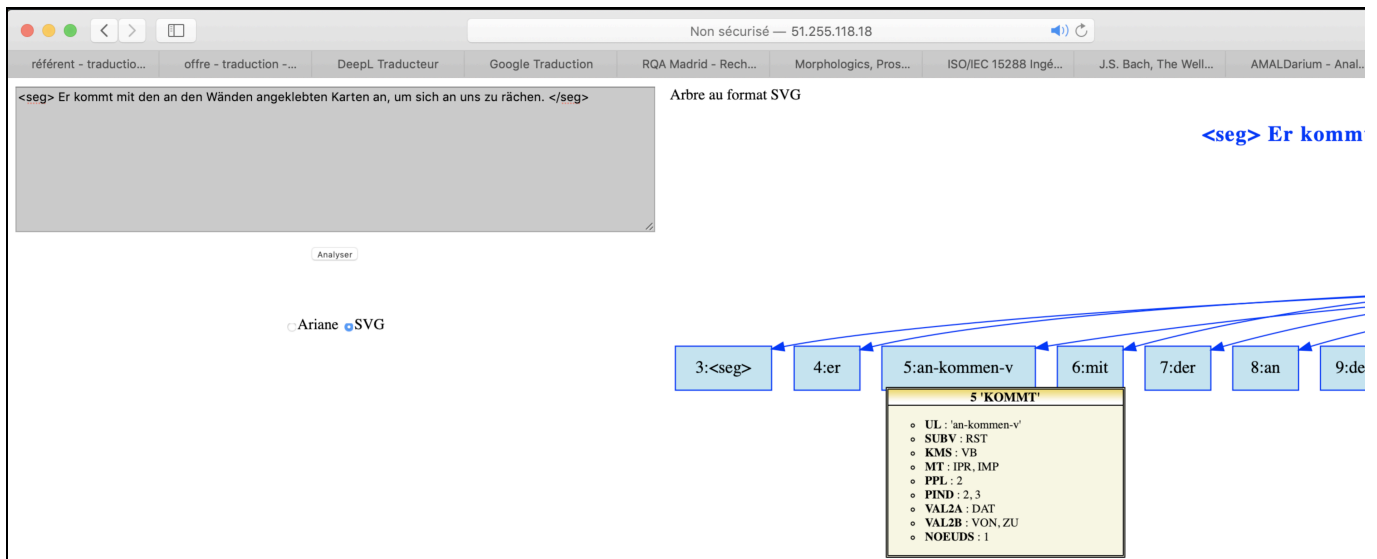
```
<txt> <!-- <txt> et pas: <txt id = "1_cle200200206"> --> <title> 1 </title>
<body> <h1> Risikoabschätzung für das familiäre Auftreten von Brustkrebs </h1> <p>
... <seg> Die Autoren legen Tabellen vor, aus denen das Erkrankungsrisiko bei
einer bestimmten familiären Vorgeschichte abhängig vom Diagnosealter von ein oder
zwei Verwandten ersten oder zweiten Grades entnommen werden kann. </seg>
```

Figure 1 : exemple de texte soumis à l'AM (XML/TEI simplifié)

¹ Plateforme générale : <http://www.lingwarium.org>, AMALD-serveur : <http://51.255.118.18/TestAnaALD>
AMALD-corpus : <http://51.255.118.18/AMALDarium>

Le résultat est un arbre décoré, visible en format texte ou graphique (SVG).

Voici une image d'écran, sur la phrase : **Er kommt mit den an den Wänden angeklebten Karten an, um sich an uns zu rächen.** On voit ici que la forme composée "kommt... an" a été reconnue, et que l'information donnée sur "kommt" est celle attachée au lemme "an-kommen".



La définition des noms et valeurs de variables (comme KMS pour catégorie morphosyntaxique, etc.) est donnée dans les « fichiers de déclarations de variables » accessibles sur le site. Notons que l'analyseur accepte en entrée un texte de longueur quelconque.

3 AMALD-corpus

Nous avons considéré plusieurs corpus médicaux en partie ou entièrement en allemand : UMLS (Browne, C., Divita, G, Aronson, R., et al., 2003), UFL medical corpus, DiK (Korte, 2019), et TIGER (Brants, S., Dipper, S., Eisenberg, P., et al., 2004). Nous n'avons pas pu les utiliser, car ils sont souvent annotés sémantiquement, dans le but de développer des techniques d'extraction de concepts, mais pas morphologiquement. Ils sont aussi prétraités, avec élimination de la casse, qui introduit des parasites dans l'AM². Il y a d'autres problèmes : par exemple (Scherrer 2008) note que TIGER n'a qu'une étiquette pour les adjectifs et les adverbes, ce qui n'est pas très satisfaisant, puisque les adjectifs se déclinent et pas les adverbes, et qu'il y a des lemmes seulement adverbiaux (sehr, zu, dann...) — alors que tout adjectif allemand peut être adverbe.

Nous sommes partis d'un corpus médical (non morphologiquement annoté) d'environ 400.000 mots concernant le cancer du sein, collecté et anonymisé vers 2013 pour étudier les différences entre deux genres, la langue de spécialité et la langue de vulgarisation, et comprenant une moitié dans chaque genre (les sources sont dans la Table 2). Le langage médical comprend beaucoup de termes composés qui sont difficiles à analyser morphologiquement, ce qui le rend très intéressant.

Nous avons extrait une partie de ce corpus et l'avons annotée morphologiquement, de façon à obtenir un premier état qu'on peut déjà considérer comme un corpus de référence. AMALD-corpus est anonymisé et vérifié. Il est pour l'instant composé de 20 fichiers (Table 1) balisés en XML/TEI (simplifié), qui sont très lisibles sous tout navigateur.

² Les noms communs prennent la majuscule à l'initiale, ce qui permet, en n+1-ième position, de distinguer par exemple 'Auftritt' (nom) de 'auftritt' (verbe).

Table 1 : 20 fichiers de AMALD-corpus

Répertoire		#occurrences textuelles	#occurrences avec balises	#occurrences de balises	% de balises	#pages standard
Fichier	1_ald	713	808	95	11,8%	2,85
Fichier	2_ald	836	966	130	13,5%	3,34
Fichier	3_ald	490	588	98	16,7%	1,96
Fichier	4_ald	525	628	103	16,4%	2,10
Fichier	5_ald	458	546	88	16,1%	1,83
Fichier	6_ald	691	815	124	15,2%	2,76
Fichier	7_ald	470	558	88	15,8%	1,88
Fichier	8_ald	731	851	120	14,1%	2,92
Fichier	9_ald	492	588	96	16,3%	1,97
Fichier	10_ald	812	961	149	15,5%	3,25
Fichier	11_ald	792	930	138	14,8%	3,17
Fichier	12_ald	901	1093	192	17,6%	3,60
Fichier	13_ald	534	618	84	13,6%	2,14
Fichier	14_ald	644	760	116	15,3%	2,58
Fichier	15_ald	1368	1602	234	14,6%	5,47
Fichier	16_ald	641	741	100	13,5%	2,56
Fichier	17_ald	742	906	164	18,1%	2,97
Fichier	18_ald	538	652	114	17,5%	2,15
Fichier	19_ald	607	725	118	16,3%	2,43
Fichier	20_ald	857	1038	181	17,4%	3,43
Total		13842	16374	2532	15,5%	55,37
Mots/page	250					

Perspective. Comme nous avons déjà soigneusement relu la totalité des 409.280 mots de ces 241 textes (de 1461 mots par texte en moyenne), il devrait être possible d'augmenter rapidement AMALD-corpus, de 20 à 240 textes, en les convertissant dans notre format d'entrée (XML/TEI simplifié) et en les soumettant à AMALD-serveur.

Références

- ALTINOK, D. *DEMorphy, German Language Morphological Analyzer*. arXiv preprint:1803.00902, 2018.
- BERMENT V., BOITET CH., GUILBAUD J.PH., KAPOCIUTE-DZIKIENE J. *Several Ways to Use the Lingvarium.org Online MT Collaborative Platform to Develop Rich Morphological Analyzers*. Computational Linguistics and Intelligent Text Processing - CICLing 2017, Revised Selected Papers, Part I, Apr 2017, Budapest, pp.81--86.
- BRANTS, S., DIPPER, S., EISENBERG, P., et al. *TIGER: Linguistic interpretation of a German corpus*. Research on language and computation, 2004, vol. 2, no 4, p. 597-620.
- BROWNE, C., DIVITA, G., ARONSON, R., et al. *UMLS language and vocabulary tools: AMIA 2003 open source expo*. In : AMIA annual symposium proceedings. American Medical Informatics Association, 2003. p. 798.
- DELPECH, E., DAILLE, B., MORIN, E. et LEMAIRE, C., *Extraction of domain-specific bilingual lexicons from comparable corpora: compositional translation and ranking*, COLING 2012, 8–12 Dec. Mumbai, 2012.
- DELPECH, E., DAILLE, B., MORIN, E. et LEMAIRE, C., *Identification of Fertile Translations in Medical Comparable Corpora: a Morpho-Compositional Approach*. In Proceedings of the 10th biennial conference of the Association for Machine Translation in the Americas (AMTA), 28 Oct.–1 Nov., San Diego, 2012.
- GUILBAUD, J.PH., BOITET, C. et BERMENT V., *Un analyseur morphologique étendu de l'allemand traitant les formes verbales à particule séparée*, TALN 2013, 17–21 juin, Les Sables d'Olonne, FRANCE, 2013.
- KORTE, L. *Online-Ressourcen zum Thema Linguistik und Medizin*. Zeitschrift für germanistische Linguistik, 2019, vol. 47, n° 1, pp. 260-268.
- LEMAIRE, C., *Un nouveau besoin dans l'industrie : une aide au « rédacteur-traduisant »*. In Actes de la 24e conférence sur le Traitement Automatique des Langues Naturelles (TALN), 26–30 juin, Orléans, 2017.
- LEMAIRE, C., GUILBAUD, J.PH., *Corpus de registres différents pour le développement d'un aligneur d'unités polylexicales*. 11e conférence sur la Lexicologie, Terminologie et Traduction (LTT), 25–28 septembre, Grenoble, France, 2018. (Article long accepté et présenté, à paraître en 2020).
- LEZIUS, W., RAPP, R. ET WETTLER, M. *A freely available morphological analyzer, disambiguator and context-sensitive lemmatizer for German*. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2. Association for Computational Linguistics, 1998. pp. 743-748.
- PISKORSKI, J., et al. *Morphisto, an open source morphological analyzer for German*. In: Finite-state Methods and NLP: Postproceedings of the 7th International Workshop FSMNLP. 2009. p. 224.
- SCHERRER, Yves. *Part-of-Speech Tagging with a Symbolic Full Parser: Using the TIGER Treebank to Evaluate Fips*. In: Proceedings of the ACL 2008 Workshop on Parsing German. 2008. pp. 16-23.

Table 2 : Sources de AMALD-corpus

Source web	Nombre d'articles	Nombre de mots
<i>aerzteblatt.de</i>	67	114406
<i>senologie.org</i>	1	81192
<i>uni-frauenklinik-</i>	7	5816
<i>iwenv.de</i>	1	1584
<i>wiralle.de</i>	1	1204
<i>krebsinformationsdienst.de</i>	1	326
<i>aerztezeitung.de</i>	1	118
<i>netdoktor.de</i>	80	45145
<i>mammamia-online.de</i>	1	39628
<i>krebsgesellschaft-nrw.de</i>	3	25697
<i>brustkrebsdeutschland.de</i>	16	20743
<i>onkosupport.de</i>	4	18559
<i>mammakarzinom-info.de</i>	36	17765
<i>mammo-programm.de</i>	6	8881
<i>de.wikipedia.org</i>	1	7759
<i>brustkrebs-sprechstunden.de</i>	1	4663
<i>ago-online.de</i>	1	4335
<i>experten-sprechstunde.de</i>	1	3820
<i>brca-netzwerk.de</i>	8	2602
<i>krebshilfe.de</i>	2	2547
<i>sportaerztebund-niedersachsen.de</i>	1	1286
<i>medical.siemens.com</i>	1	1204
TOTAL	241	409280

Générateur de dictionnaires au format Android pour les langues peu dotées

Rémy Bonnet¹ Guillaume Segerer¹

(1) LLACAN, 7 rue Guy Môquet - BP 8, 94801 Villejuif, France

remy.bonnet@cncrs.fr, guillaume.segerer@cncrs.fr

RÉSUMÉ

Nous présentons un outil informatique en source libre permettant, à partir d'un dictionnaire sous format électronique de générer une application Android. L'objectif est de tirer profit de l'augmentation spectaculaire des utilisateurs de mobiles multifonctions dans de nombreuses régions du monde, notamment en Afrique, pour permettre aux locuteurs de langues moins dotées de bénéficier de dictionnaires de bonne qualité, faciles à utiliser et de s'impliquer dans la collecte de données. Cet outil est conçu pour être compatible avec les principaux formats lexicographiques.

ABSTRACT

Dictionary App Generator for Less Resourced Languages.

This article presents an open source dictionary Android app generator project for less resourced languages. If these dictionaries can be useful to linguists our main goal is to take advantage of the dramatic increase in the number of smart-phone owners in many parts of the world and especially Africa in order to allow native speakers to benefit from an easily accessible and good quality version of a dictionary of their language and interact with them linguists who work on it. Our application is meant to adapt to as many formats as possible.

MOTS-CLÉS : Dictionnaire, Application Android, Langues peu dotées.

KEYWORDS: Dictionary, Android Application, Less Resourced Languages

1 Introduction

Ces dernières années les linguistes de terrain ont pu, grâce aux avancées des technologies numériques, archiver numériquement leur données et ajouter une composante multi-media au traditionnel triptyque "grammaires, dictionnaires, et recueils de textes" (Musgrave & Thieberger, 2014).

Parallèlement l'usage des téléphones mobiles s'est considérablement répandu et démocratisé dans des endroits du monde où l'usage des ordinateurs reste assez marginal. Ainsi le nombre de souscripteurs à des forfaits internet mobiles a quadruplé lors de la dernière décennie (Poushter). L'immense majorité des téléphones mobiles vendus aujourd'hui sont des mobiles multifonctions.

Le système d'exploitation mobile Android, lancé en 2005 par Google équipe aujourd'hui plus de 80 % des mobiles multifonction ([Auffray 2016](#)).

Une application Android semble donc être, pour de nombreuses communautés linguistiques minoritaires, un moyen d'accéder à un dictionnaire qui, outre son intérêt linguistique, est souvent perçu comme un résultat tangible du travail des linguistes et peut constituer une incitation à s'impliquer dans la collecte des données d'autant que les mobiles multifonctions, en tant qu'outils de communication, permettent aux utilisateurs de contacter directement les auteurs.

L'outil présenté ici, dont le développement a été financé par le projet DReaM ([site internet](#)) et fait suite à une version embryonnaire développée pour le LabEx EFL sous la supervision de Guillaume Jacques et Alexis Michaud, a pour objectif de permettre aux linguistes ayant un dictionnaire numérique de générer une application Android.

2 Application Android

Avant d'évoquer le générateur d'applications dictionnaire lui-même, il convient de présenter le type d'application qu'il génère. Le code étant ouvert les créateurs d'applications ont tout loisir de modifier le rendu final selon leurs besoins.

L'application est constituée de différentes pages :

- Page de présentation : une image qui apparaît 5s. au lancement de l'application.
- Page principale (*fig. 1*) : cette page permet d'accéder directement à une entrée en commençant à la taper et en choisissant parmi les propositions automatiques. Il est aussi possible de lancer une recherche à partir d'une définition dans une des langues de traduction ou d'accéder à la page de recherches avancées ou simplement de faire afficher tous les entrées commençant par une lettre.
- Page de recherche avancée : permet de faire une recherche plus fine avec divers filtres et conditions.
- Page liste de mots : affiche la liste des résultats d'une recherche faite à partir d'une traduction, d'un thésaurus etc.
- Page de description d'une entrée (*fig. 2*) : Cette page présente une entrée particulière, on y trouve les informations classiques (partie du discours, définitions dans les langues choisies par l'auteur du dictionnaire...) des informations complémentaires (thésaurus, notes anthropologiques...) des mises en contextes (aussi bien des exemples classiques que des expressions figées, des proverbes...) des liens vers d'autres entrées (synonymes, hyperonymes...) ainsi que des fichiers son ou image.
- Pages techniques : informations sur le dictionnaire, sur l'application, préférences...



Figure 1: Page principale

- Page d’interaction avec l’utilisateur : si il est possible d’envoyer un mail à partir de la page principale, il existe une page spécifique accessible depuis la page de description d’une entrée. Elle permet d’envoyer un commentaire au sujet d’une entrée spécifique. Il est possible de saisir le commentaire qui apparaîtra dans le corps du mail mais aussi d’enregistrer un fichier son qui sera ajouté en pièce jointe.

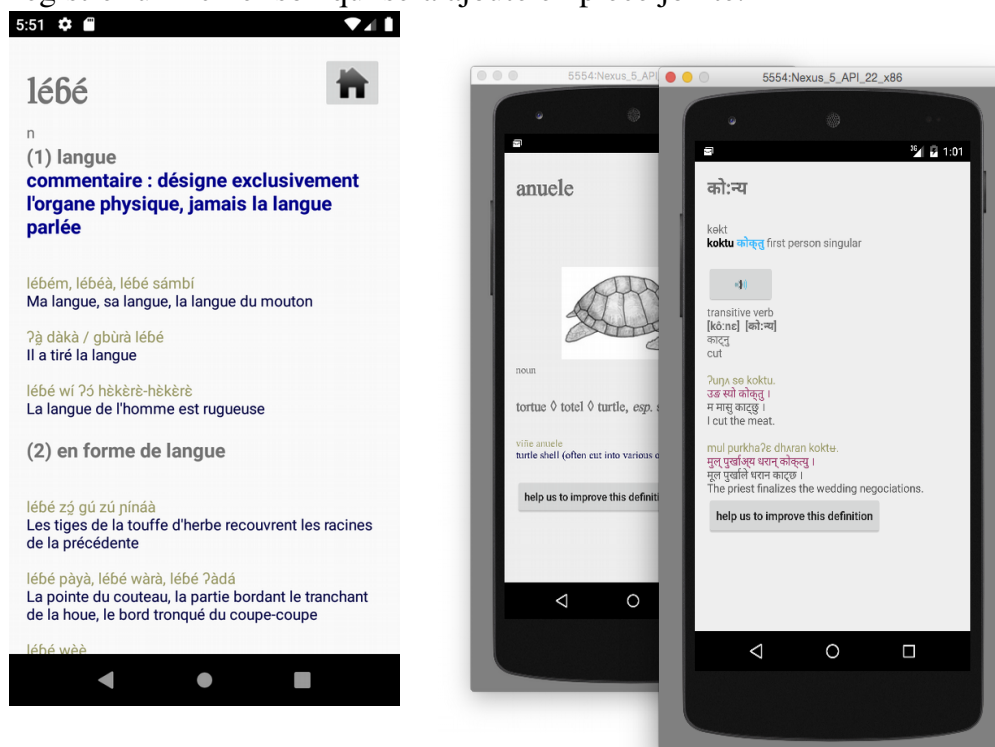


FIGURE 2: Exemples de page description

3 Générateur d’application

Le générateur d’applications peut être utilisé avec une ligne de commande ou à l’aide d’une petite interface (fig. 3). Conformément à la philosophie du logiciel libre, notre outil produit un code source. Ce code est écrit en Java (licence publique générale GNU). Java a été choisi car il est l’un des deux langages de programmation officiellement pris en charge par Android et qu’il reste un des langages informatiques les plus utilisés. Le code s’accompagne d’une base de données (bdd) en SQLite dans laquelle sont stockées les données du dictionnaire. Le format de la bdd est inspiré de LMF (Francopoulo, 2013) et compatible avec les formats lexicaux les plus utilisés (notamment MDF et LIFT qui doivent leur succès aux applications Toolbox et Flex développées par la SIL). Pour mettre à jour une application suite à une modification du dictionnaire il suffit de mettre à jour la bdd. Pour générer la base de données il faut fournir à l’application, outre le dictionnaire lui-même, un fichier décrivant les balises spécifiques à ce dictionnaire, l’ordre alphabétique de la langue (il est possible de spécifier les caractères spéciaux ou les lettres ne pouvant pas être à l’initiale d’un mot) et l’emplacement des éventuels fichiers son et image. Le générateur a donc un mode création d’application et un mode mise à jour qui se contente de modifier la bdd.

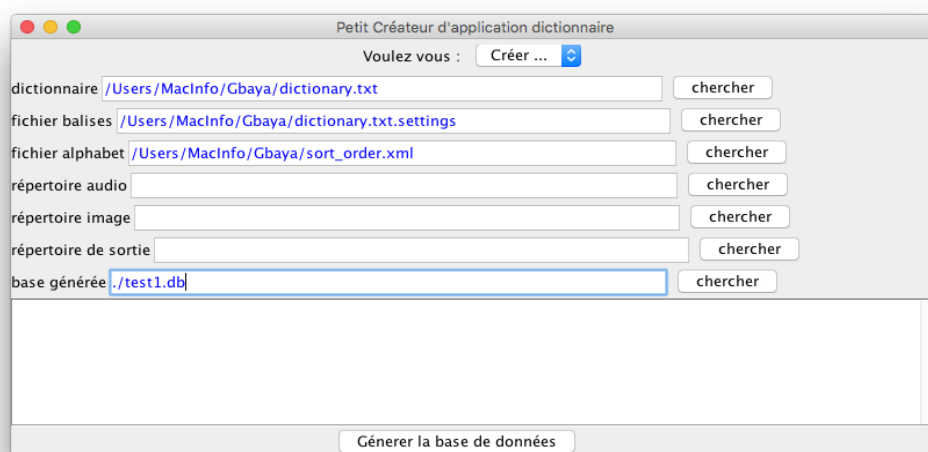


FIGURE 3: Page principale de l'outil de génération

Le code source est peu différent d'une application à une autre, c'est pourquoi sa génération nécessite seulement les informations supplémentaires par rapport à la création d'une bdd : les noms du ou des auteurs, l'adresse électronique à laquelle seront envoyées les commentaires des utilisateurs, l'image à afficher dans la page de présentation ainsi que l'image à utiliser pour l'icône de l'application (qui apparaît sur le menu du terminal mobile).

4 Premiers résultats

Outre trois dictionnaires (khaling, japhug et bena) actuellement accessibles mais issus d'une version antérieure de l'application, la version électronique du dictionnaire gbaya écrit par Paulette Roulon Doko est actuellement en version beta sur le "Play Store" de Google, elle sera accessible au grand public dès la fin de la période de confinement. Des dictionnaires kakabé, bambara, wolof et keeraak seront tester par des utilisateurs volontaires à ce moment là. D'autres dictionnaires devraient suivre rapidement notamment des dictionnaires plus anciens comme le dictionnaire shisumbwa-français écrit en 1901 par le père Capus qui a été numérisé. Le code source du générateur d'applications dictionnaire sera disponible sous licence open-source d'ici la fin de l'été 2020 sur le site du LLACAN (<http://llacan.vjf.cnrs.fr/>).

Références

- AUFFRAY C. (2016) *Chiffres clés : les OS pour smartphones* ([zdnet](#))
- FRANCOPOULO G. (Ed.). (2013). *LMF: Lexical Markup Framework*. Wiley Online Library.
- MUSGRAVE S. & THIEBERGER N. (2014, November). *Rethinking grammatical description: from Heath to hypertext*. Lecture, Research Unit for Indigenous Language, University of Melbourne.
- POUSHTER J. (2016). Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies in <https://www.pewresearch.org/global/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/>

Désidentification de données texte produites dans un cadre de relation client

Guillaume Dubuisson Duplessis Elliot Bartholme Sofiane Kerroua
Mathilde Poulain Ahès Roulier Anne-Laure Guénet

EDF Commerce, Direction Numérique, Tour PB6, 178 Rond-Point de la Défense, 92800 Puteaux, France
{guillaume.dubuisson-duplessis, mohamed-sofiane.kerroua,
anne-laure.guenet}@edf.fr

RÉSUMÉ

Cette démonstration présente une solution performante de désidentification de données texte selon 13 types d'entités nommées et entraînée sur des données issues de la relation client.

ABSTRACT

De-identification of customer relationship text data

This demonstration presents an efficient text de-identification system based on 13 named-entity types. It has been trained on customer relationship data.

MOTS-CLÉS : désidentification, reconnaissance d'entités nommées, RGPD.

KEYWORDS: de-identification, named-entity recognition, GDPR.

1 Désidentification de données texte à EDF Commerce

1.1 Minimiser l'usage des données à caractère personnel

Chaque mois des millions de données texte sont produites dans le cadre de la relation client au sein d'EDF Commerce de la part des clients (e.g., e-mails, réponses libres à des questionnaires de satisfaction) et des conseillers (e.g., commentaires de contact). Ces données majoritairement en français sont riches : elles offrent un large panel de structures allant d'expressions libres et spontanées à des formes contraintes comme des formulaires ; tout en manifestant une grande variabilité en termes de respect de l'orthographe, de la syntaxe et de niveau de langue. Ces données sont utilisées pour répondre au mieux aux attentes de nos clients. En outre, elles sont exploitées dans de nombreux cas d'utilisation « métier » visant à optimiser la relation client. Le cas le plus répandu est celui des tableaux de bord permettant aux opérationnels de suivre une activité comme, par exemple, le traitement des e-mails client (Dubuisson Duplessis *et al.*, 2019). Afin de nourrir ces cas d'utilisation, des données texte peuvent être consultées par des individus (e.g., lors de tâches d'annotation manuelle, lors de retour aux données dans des tableaux de bord), et intervenir dans de nombreuses tâches de modélisation du TALN (e.g., classification, analyse de ressenti client). Or, ces données texte font généralement apparaître des données à caractère personnel (DCP) telles que les noms, prénoms, adresses postales, e-mails, et de nombreux identifiants numériques. Le règlement général sur la protection des données (RGPD) renforce les droits des individus quant à l'utilisation qui peut être

faite de leurs DCP. A cette fin, il exige une minimisation de l’usage des DCP au regard de la finalité pour laquelle elles sont traitées.

1.2 Une approche par désidentification pour respecter les contraintes du RGPD

Afin de limiter efficacement l’usage des DCP, notre approche vise à les supprimer par une procédure de désidentification automatique qui fonctionne en deux temps : (i) une phase de reconnaissance d’entités nommées (NER) correspondant aux DCP (Nouvel *et al.*, 2016), suivie de (ii) une phase de délexicalisation substituant le texte des entités nommées par le type des entités nommées. Par exemple, « Je suis Jean Dupont (Paris 12^e). Je ne comprends pas ma facture n° 12 345 6. » peut être délexicalisé en « Je suis _PERSON_ (_LOCALISATION_). Je ne comprends pas ma facture n° _NUMBER_. ». L’avantage de la désidentification est de prévenir la divulgation de DCP à des personnes non-habilitées tout en conservant la substance du document. Dans le cadre de la modélisation, la désidentification en amont sur les données utilisées permet de prévenir des biais d’apprentissage en empêchant l’utilisation de DCP dans les décisions des algorithmes. Par exemple, elle évite l’impact du genre ou de l’origine des noms et prénoms. Une limite notable d’une approche par désidentification est atteinte par l’usage de périphrase (« Président de la République française, je souhaite . . . »). L’évaluation de nos algorithmes indique que ce type de phénomène est très rare dans nos données de relation client.

1.3 Une solution de désidentification performante en français

La construction de notre solution de désidentification a bénéficié d’un processus d’annotation rigoureux et de qualité. 13 types d’entités nommées parmi lesquels des adresses, des noms/prénoms, des e-mails, des informations bancaires et des numéros client ont été annotés en interne via une plateforme web inspirée du projet Camomile (Poignant *et al.*, 2016). Le corpus d’apprentissage contient environ 1000 e-mails et 700 conversations de chat client/conseiller représentant 11459 tours de parole. Ces données sont en français. Le corpus contient approximativement 6200 instances d’entités. La répartition des catégories d’entités est détaillée dans le Tableau 1. Les e-mails et les conversations

Localisation	Personne	Info. numériques	URL/e-mails	Info. bancaire
19.1% / 41.9%	47.7% / 27%	29.6% / 22.9%	3.3% / 8%	< 0.5%

TABLE 1 – Répartition des catégories d’entités dans le corpus d’apprentissage. Lecture : proportion en nombre d’instances d’entité / proportion en nombre de tokens.

sont pré-traités pour uniformiser l’encodage (en particulier au niveau des « smileys »), supprimer les balises et normaliser les entités HTML. Les données de conversations de chat ont été simplement segmentées au niveau des tours de parole. Les e-mails ont subi une segmentation plus lourde visant à isoler les informations techniques (par exemple, les informations d’en-têtes telles que l’expéditeur, les destinataires, les dates), des champs structurés (par exemple, « Numéro client : . . . » et des champs de texte libre. Les informations techniques et les champs structurés sont découpés via des heuristiques. Les parties de texte libre sont segmentées en phrases.

Algorithmiquement, notre solution se fonde sur l’hybridation de règles et d’apprentissage profond afin d’obtenir le meilleur compromis entre performance et temps de calcul. L’approche par règles est

utilisée pour les entités fortement structurées et peu dépendantes du contexte comme les identifiants numériques, les informations bancaires, les e-mails et les URL. L’approche par apprentissage profond se concentre sur les entités telles que les prénoms, noms, adresses postales et lieux. Plusieurs modèles sont maintenus afin de maximiser la performance en NER tout en satisfaisant les deux principales contraintes techniques impliquées que sont le temps de calcul et les limitations liées à la taille des documents. La première partie de nos modèles se fonde sur la famille des RNN et utilise des bi-LSTM+CRF (Akbik *et al.*, 2018) combinant des plongements vectoriels au niveau des caractères (Lample *et al.*, 2016), des sous-tokens (Heinzerling & Strube, 2018), et des tokens (Pennington *et al.*, 2014). La seconde partie se fonde sur l’architecture « transformer » (Vaswani *et al.*, 2017) et utilise fructueusement les modèles en français créés par la communauté tels que CamemBERT (Martin *et al.*, 2020) et FlauBERT (Le *et al.*, 2020).

De par sa fondation sur des données variées de la relation client, notre système peut être considéré comme spécialisé pour ce domaine. Néanmoins, les types d’entités traités ont une portée générique à l’exception de quelques identifiants numériques spécifiques à EDF Commerce (par ex., numéro client, numéro de contrat, index de compteur).

Notre solution obtient de bonnes performances en NER sur des énoncés de chat conseiller/client confirmant ainsi la maturité des approches par apprentissage profond pour des cas opérationnels (Barriere & Fouret, 2019). Sur les onze entités gérées par la partie « règle », nous obtenons une micro F1 à 0.94 sur notre test. Sur les deux entités gérées par la partie « apprentissage profond », nous obtenons une micro F1 à 0.97. En outre, nous avons évalué manuellement la performance en anonymisation sur 1000 conversations. Nous avons lu des conversations désidentifiées en cherchant à recroiser des informations dans le texte pour vérifier leur anonymisation. Sur cette évaluation exigeante, nous avons obtenu un taux d’anonymisation de 96%. Cette évaluation nous a permis d’identifier quelques cas limites pour notre système. Sans prétendre à une présentation exhaustive, notre système semble manquer certaines entités peu communes (comme des noms de village, des noms de personne ou des prénoms peu courants) ou à l’opposé des mots plutôt communs (par exemple, un nom de famille pouvant dénoter un métier comme « M. Boulanger », un prénom comme « Claire » qui peut également être un adjectif). Les erreurs de typographie, variées dans des expressions spontanées de client, fragilisent également la reconnaissance des entités nommées. Sur les conversations client-conseiller, cela s’est particulièrement manifesté autour de la reconnaissance d’adresses e-mails. Une perspective intéressante de nos travaux réside en une analyse quantitative et qualitative approfondie des erreurs systématiques de notre solution.

2 Démonstrateur

Notre solution de désidentification est destinée à être utilisée aussi bien en amont des projets « data » afin de désidentifier les données textes avant qu’elles ne soient manipulées qu’en production dans projets industrialisés. Un cas d’utilisation fréquent est celui des tableaux de bord de pilotage permettant des retours aux données anonymisées (voir, par exemple, le projet Cameli@ (Dubuisson Duplessis *et al.*, 2019)).

Cette démonstration propose une interface permettant de saisir un texte libre en se mettant à la place d’un client EDF et de le désidentifier suivant 13 types d’entités nommées en utilisant les algorithmes précédemment évoqués. Le démonstrateur inclut la possibilité de tester plusieurs approches alternatives en laissant libre choix du système d’apprentissage profond utilisé (RNN, « transformer »).

Remerciements

Nous remercions chaleureusement toutes les personnes qui sont intervenues de près ou de loin sur ce projet : Gilles Pouëssel, Mélanie Cazes, Meryl Bothua, Lou Charaudeau, Ibtissem Menacer, Uta Hosokawa, Aurore Hamimi, Anaël Cabrol, Sylvain Boucault, François Bullier, Jean Vidal et Marie Hervé.

Références

- AKBIK A., BLYTHE D. & VOLLGRAF R. (2018). Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, p. 1638–1649.
- BARRIERE V. & FOURET A. (2019). May I Check Again ? A simple but efficient way to generate and use contextual dictionaries for Named Entity Recognition. Application to French Legal Texts. arXiv preprint : [1909.03453](https://arxiv.org/abs/1909.03453).
- DUBUISSON DUPLESSIS G., KERROUA S., KUZNIK L. & GUÉNET A.-L. (2019). Cameli@ : analyses automatiques d’e-mails pour améliorer la relation client. *Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, p. 623–625.
- HEINZERLING B. & STRUBE M. (2018). BPEmb : Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*.
- LAMPLE G., BALLESTEROS M., SUBRAMANIAN S., KAWAKAMI K. & DYER C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 260–270, San Diego, California : Association for Computational Linguistics. DOI : [10.18653/v1/N16-1030](https://doi.org/10.18653/v1/N16-1030).
- LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). Flaubert : Unsupervised language model pre-training for french. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*. arXiv preprint : [1912.05372](https://arxiv.org/abs/1912.05372).
- MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. arXiv preprint : [1911.03894](https://arxiv.org/abs/1911.03894).
- NOUVEL D., EHRLMANN M. & ROSSET S. (2016). *Named Entities for Computational Linguistics*. John Wiley & Sons.
- PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543.
- POIGNANT J., BUDNIK M., BREDIN H., BARRAS C., STEFAS M., BRUNEAU P., ADDA G., BESACIER L., EKENEL H., FRANCOPOULO G. *et al.* (2016). The Camomile collaborative annotation platform for multi-modal, multi-lingual and multi-media documents. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.
- VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. In *Advances in neural information processing systems*, p. 5998–6008.

Analyse sémantique de transcriptions automatiques d'appels téléphoniques en français

Emmanuelle Dusserre Ruslan Kalitvianski Mathieu Ruhlmann Muntsa Padró
Eloquent, 5 allée de Palestine, 38610 Gières, France
{prenom.nom}@eloquant.com

RÉSUMÉ

Dans cet article, nous présentons la mise en œuvre d'une chaîne de traitement sémantique complète dédiée aux conversations audio issues de centres d'appel téléphoniques, depuis la phase de transcription automatique jusqu'à l'exploitation des résultats, en passant par l'étape d'analyse sémantique des énoncés. Nous décrivons ici le fonctionnement des différentes analyses que notre équipe développe, ainsi que la plateforme interactive permettant de restituer les résultats agrégés de toutes les conversations analysées.

ABSTRACT

Semantic analysis of automatic phone call transcriptions in French

In this article, we present the implementation of a complete semantic processing chain dedicated to call center phone conversations, from the speech-to-text phase to the exploitation of the results, including the semantic analysis of the utterances. Here we describe the workings of the various analyses that our team develops, as well as the interactive platform that displays the aggregated results of all the analyzed conversations.

MOTS-CLÉS : analyse sémantique, centres d'appels, transcriptions automatiques, *speech-to-text*
KEYWORDS: semantic analysis, speech-to-text, call centers

1 Introduction

Notre société propose une gamme complète de logiciels pour la gestion de la relation client. Nous proposons, entre autres, des logiciels pour la gestion des centres d'appel, pour la réalisation d'enquêtes de satisfaction, ainsi que des modules pour faire l'analyse sémantique du texte. L'un des projets les plus innovants de notre entreprise est le développement des modules d'analyse sémantique des transcriptions automatiques de conversations téléphoniques, modules qui sont clés pour une meilleure exploitation des données audios collectées par nos clients.

Cet article décrit, d'une part, la plateforme en ligne qui restitue de manière graphique et interactive les résultats des analyses sémantiques quotidiennes et, d'autre part, le fonctionnement du moteur d'analyse sémantique sous-jacent et de ses modules. L'analyse s'appuie sur des transcriptions automatiques réalisées par un partenaire, qui effectue également une diarisation.

La restitution présentée dans cette démonstration est destinée aux chefs de projet des services de consommateurs de notre client, qui désirent avoir un panorama synthétique et interactif des interactions téléphoniques entre les agents de support de leur entreprise et leurs clients.

2 L'analyse sémantique : briques et mode opératoire

Nous traitons de manière quotidienne les audios transmises par notre client de façon sécurisée, datant de la veille. Les audios sont au format `wav`, en monocanal, 8 KHz, 64 Kbit/seconde, encodées en G.711. Ces données sont accompagnées d'un fichier de métadonnées qui renseigne des informations utiles au client au moment de la restitution des analyses (date et heure, pôle, etc.).

Les audios sont envoyées à la transcription, réalisée par un partenaire, qui retourne des documents XML contenant, d'une part une information sur le nombre et le genre des locuteurs, et d'autre part, la transcription segmentée en mots chronodatés, regroupés en énoncés étiquetés par l'identifiant de leur locuteur. Ici les énoncés sont des segments de parole délimités soit par des silences, soit par les énoncés d'un autre locuteur.

Ces transcriptions et les métadonnées des audios sont converties vers la structure de données unique sur laquelle opèrent l'ensemble de nos modules d'analyse sémantique, qui sont ordonnés en chaîne de traitement *pipeline*. Chaque conversation est analysée de façon indépendante des autres conversations, et l'unité d'analyse est l'énoncé dans la plupart des cas.

Nos modules d'analyse sémantique s'appuient sur une analyse morphosyntaxique réalisée avec Talismane (Urieli & Tanguy, 2013), et incluent les analyses sémantiques suivantes, réalisées dans cet ordre :

- Une classification des locuteurs (identifiés en amont par la diarisation), en « agent » ou « client ». Cela facilite la lecture de la transcription et permet de filtrer les énoncés dans l'interface de restitution décrite dans la section 3. Cette classification est effectuée sur les ensembles des énoncés produits par chaque locuteur. Elle est réalisée par un classifieur binaire MaxEnt (Nigam et al., 1999) entraîné avec des n-grammes extraits de 200 conversations annotées manuellement et choisies sur la base de leurs durées (afin de capter les différences lexicales observées entre les conversations longues, moyennes et courtes).
- Une anonymisation des noms, lieux, numéros (tout en évitant d'anonymiser les numéros de références produit), adresses physiques et électroniques. Cela garantit un niveau de confidentialité qui nous permet contractuellement de conserver les transcriptions analysées pour une durée d'un an. Cette anonymisation est réalisée à l'aide de listes de prénoms, noms, noms de villes et régions, et de plusieurs dizaines d'expressions régulières étalonnées sur environ 1200 énoncés contenant diverses informations personnelles.
- Une extraction de concepts récurrents : il s'agit de noms ou groupes nominaux significatifs extraits selon le modèle proposé par (Sclano & Velardi, 2007) qui compare les fréquences d'apparition des termes entre le corpus spécifique et un corpus généraliste.
- Une extraction d'entités nommées du domaine de notre client, fondée sur une liste de termes établis avec le client et de leurs synonymes. Nous avons constitué différents gazetteers, spécifiques aux entités nommées du secteur de notre client, afin de relever toutes les occurrences de ces termes dans les conversations. Les gazetteers suivent une syntaxe relativement simple : le terme, qui peut s'écrire sous forme d'expression régulière, et le tag sémantique qu'on souhaite lui attribuer. Par exemple, le nom d'un produit a pour tag sémantique la famille du produit à laquelle il se réfère.
- Un étiquetage multi-label de chaque énoncé par un sous-ensemble d'une vingtaine de catégories pertinentes pour le métier de notre client, elles-mêmes groupées en une dizaine de méta-catégories. Ce traitement est fondé sur l'apprentissage automatique (utilisant comme traits les mots, leurs lemmes, les groupes nominaux, et les annotations sémantiques réalisées en amont via notamment des listes de gazetteers propres au domaine), et sur des règles expertes exprimées

à l'aide du formalisme TokensRegex (Chang & Manning, 2014). Ces dernières nous permettent d'une part d'enrichir sémantiquement les données destinées à la classification en annotant les expressions polylexicales et, d'autre part, de corriger les catégories là où l'apprentissage machine rencontre des difficultés. Pour ce second cas, les TokensRegex ajustent les scores attribués par le classifieur lorsqu'elles rencontrent une expression polylexicale, en revalorisant le score sur la catégorie à laquelle l'expression appartient.

- Un étiquetage multi-label de la conversation complète, qui agrège d'une façon particulière l'ensemble des étiquettes attribuées aux énoncés de la conversation, en suivant les règles de dominance thématique spécifiées par notre client.
- Une extraction des fragments exprimant une opinion positive ou négative. Elle s'appuie sur des règles expertes TokensRegex développées spécifiquement pour les audios, car nous nous sommes aperçus que nos grammaires d'opinions développées pour les verbatims écrits étaient inadaptées pour les transcriptions de conversations : outre les erreurs de transcriptions, le langage parlé diffère fortement du langage écrit, que ce soit par sa structure rhétorique et discursive ou par le registre et les choix lexicaux.

La *pipeline* produit en sortie, pour chaque conversation, une sérialisation XML de l'ensemble des énoncés enrichis de leurs métadonnées et de leurs analyses. Ces fichiers XML sont ensuite indexés sur un serveur Elasticsearch¹ et restitués dans une instance de la plateforme interactive Kibana², décrite ci-dessous.

3 L'interface exploratoire des résultats

L'interface web de la plateforme restitue d'une part les métadonnées des audios (volumétrie, durée moyenne, source) et, d'autre part, les résultats de l'analyse sémantique via un tableau de bord qui inclut une quinzaine de visualisations interactives sous forme d'histogrammes, de « camemberts » ou de nuages de mots (cf. Figure 1), groupées en cinq sections : « Métadonnées », « Concepts et Produits », « Catégorisation », « Opinions » et « Conversations ». En cliquant sur les éléments d'une visualisation un filtre s'applique pour mettre à jour l'ensemble du tableau de bord.

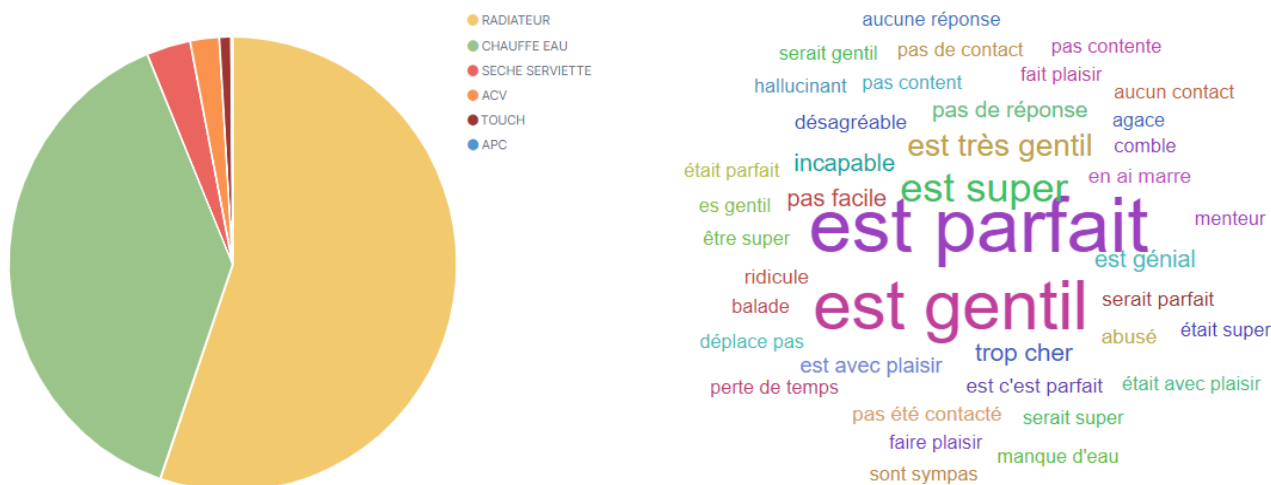


FIGURE 1 : illustration d'un « camembert » des pôles produits abordés dans les conversations, et d'une détection de fragments portant une polarité positive ou négative.

¹ <https://www.elastic.co/fr/elasticsearch>

² <https://www.elastic.co/fr/kibana>

Les visualisations servent à afficher de façon graphique la fréquence et/ou la proportion d'éléments relevés. Ainsi la taille des mots ou des expressions dans les nuages de mots correspondent à leurs fréquences relatives.

La Figure 2 ci-dessous illustre un prototype d'affichage des conversations transcrites et analysées.



FIGURE 2 : illustration du prototype du lecteur audio synchronisé avec la transcription analysée. Les pastilles colorées et la coloration des énoncés représentent les catégories attribuées aux énoncés et leurs occurrences dans la conversation. Actuellement cet affichage n'illustre pas la détection des opinions, ni l'aspect multi-label de la catégorisation.

Remerciements

Nous voudrions remercier David Graceffa, Tim Robert, Mathieu Bliem, Daniel Chipan et Hui Pillot pour leur travail sur le prototype du lecteur audio présenté ci-dessus, ainsi que Myriam El Helou pour sa participation aux modules d'analyse sémantique. Nous remercions également le comité de relecture pour leurs critiques constructives.

Références

- CHANG A. X., & MANNING C. D. (2014). TokensRegex: Defining cascaded regular expressions over tokens. *Stanford University Computer Science Technical Reports. CSTR 2*: 2014.
- NIGAM K., LAFFERTY J., & MCCALLUM A. (1999). Using maximum entropy for text classification. *In IJCAI-99 workshop on machine learning for information filtering (p. 61-67)*. 1999.
- SCLANO, F., ET VELARDI, P. (2007). TermExtractor: a Web Application to Learn the Shared Terminology of Emergent Web Communities. *In R. J. Gonçalves, J. P. Müller, K. Mertins, et M. Zelm (Éd.), Enterprise Interoperability II (p. 287-290)*, 2007. Springer London.
- URIELI A. & TANGUY L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talisman. *20e conférence du Traitement Automatique du Langage Naturel (TALN)*, 2013, Sables d'Olonne, France.

Muzeeglot : annotation multilingue et multi-sources d'entités musicales à partir de représentations de genres musicaux

Elena V. Epure Guillaume Salha Félix Voituret
Marion Baranes Romain Hennequin
Deezer Research, Paris, France
research@deezer.com

RÉSUMÉ

Au sein de cette démonstration, nous présentons Muzeeglot, une interface web permettant de visualiser des espaces de représentations de genres musicaux provenant de sources variées et de langues différentes. Nous montrons l'efficacité de notre système à prédire automatiquement les genres correspondant à une entité musicale (titre, artiste, album...) selon une certaine source ou langue, étant données des annotations provenant de sources ou de langues différentes.

ABSTRACT

Muzeeglot : cross-lingual multi-source music item annotation from music genre embeddings

In this demonstration, we present Muzeeglot, a web interface providing a visualization of multi-sources and multilingual music genres embedding spaces. We demonstrate the ability of our system to automatically infer the genres annotations of a music entity (track, artist, album...) according to some source or language, based on annotations from different sources or languages.

MOTS-CLÉS : Genres Musicaux Multilingues, Annotation Automatique d'Entités Musicales, Espaces de Représentations Sémantiques.

KEYWORDS: Cross-lingual Music Genres, Music Entity Annotation, Semantic Embedding Spaces.

1 Introduction

Pour les plateformes de streaming musical, l'exploitation des données liées aux genres musicaux est essentielle (Mandel *et al.*, 2010; Schedl & Ferwerda, 2017). Ces genres peuvent en effet refléter et résumer les préférences musicales des utilisateurs, permettre d'améliorer l'organisation d'un catalogue musical ou encore être utilisés par des systèmes de recommandation pour proposer du contenu musical personnalisé. Toutefois, relier des genres musicaux à un titre ou à un artiste est une tâche difficile, en raison de la nature subjective de ces concepts. Des études ont montré qu'il existe une grande disparité dans la façon dont les gens associent les genres à la musique, selon leurs préférences, leurs connaissances musicologiques et leur culture (Sordo *et al.*, 2008; Lee & Downie, 2013).

Les plateformes de streaming musical collectent des annotations d'entités musicales (titres, artistes, albums...) par genre auprès de nombreux fournisseurs de musique, provenant du monde entier. Homogénéiser ces différentes annotations est crucial pour l'analyse de données et la recommandation, mais est particulièrement difficile (Hennequin *et al.*, 2018; Epure *et al.*, 2019). Outre les disparités culturelles et musicologiques mentionnées précédemment, des différences d'annotations entre les

sources peuvent également apparaître suite à l'utilisation de niveaux de détails inégaux (selon la spécialisation de la source), à un manque d'exhaustivité (bien qu'il existe des milliers de genres, certaines sources n'en utilisent qu'un petit sous-ensemble pour l'annotation) ou encore suite à des incohérences orthographiques entre les différentes sources (certains genres sont annotés sous des formes différents : *d&b* vs. *drum and bass*).

Au sein de cette démonstration, nous présentons Muzeeglot, un système qui, étant données les annotations d'entités musicales par genres selon de multiples sources (ces sources étant désignées comme des *taxonomies* par la suite) pouvant être de langues différentes, prédit automatiquement les annotations équivalentes au sein d'une taxonomie cible. La solution proposée est non-supervisée, et est fondée sur l'apprentissage d'espaces de représentations sémantiques et multilingues des genres musicaux. Muzeeglot est accessible publiquement sur la page GitHub de Deezer¹.

2 Solution et démonstration

Afin d'aboutir à de tels espaces de représentations, nous calculons tout d'abord les « *word embeddings* » des genres musicaux apparaissant dans les taxonomies sources et dans la taxonomie cible. Il s'agit de représentations vectorielles de mots dans un espace de faible dimension, apprises à partir de leurs cooccurrences au sein de grands corpus textuels (Pennington *et al.*, 2014; Grave *et al.*, 2018). Dans cette démonstration, nous utilisons la version multilingue de fastText (Mikolov *et al.*, 2018; Joulin *et al.*, 2018) pour la construction de ces représentations. Nous les consolidons afin qu'elles correspondent davantage à notre domaine d'application : la musique. En effet, selon fastText, *house* est plus proche de *building* que de *music*, et *jazz* est plus proche de *folk* que de *bepop*. Pour pallier ces ambiguïtés, liées à l'utilisation de corpus d'entraînement très généraux, nous créons un graphe de connaissances des genres musicaux à partir de DBpedia (Auer *et al.*, 2007). Ce graphe relie les genres, de différentes langues, selon plusieurs types de relations (sous-genre, genre dérivé, fusion...). Via l'algorithme de *retrofitting* (Faruqui *et al.*, 2015), nous ajustons les représentations des genres musicaux en y incorporant l'information sémantique du graphe. Les vecteurs des genres musicaux connectés dans le graphe auront ainsi tendance à être plus proches dans l'espace final.

Muzeeglot est une interface web, présentant une visualisation en 3D de cet espace de représentations des genres musicaux. Les Figures 1 et 2 proposent un aperçu du rendu final. Le frontend de Muzeeglot est basé sur le framework VueJS, des composants UX Vuetify, et Plotly pour le rendu 3D avec WebGL. Le backend est développé en Python3 via FastAPI, et les données sont stockées dans un datastore Redis. Sur cette interface web, l'utilisateur a la possibilité de sélectionner :

- une entité musicale, tel que le groupe Pink Floyd au sein des Figures 1 et 2 ;
- plusieurs taxonomies sources, pour lesquelles les genres correspondants s'afficheront ;
- une taxonomie cible, pour lesquels Muzeeglot devra prédire les annotations équivalentes de genres pour l'entité musicale sélectionnée.

Un Top 10 des prédictions des genres musicaux de cette entité au sein de la taxonomie cible sera alors proposé (voir Figure 2). Ces prédictions seront calculées à partir de la *similarité cosinus* moyenne entre les représentations des genres de la taxonomie cible et les représentations des genres de l'entité musicale sélectionnée au sein des taxonomies sources. Les dix genres de la taxonomie cible ayant la plus forte similarité cosinus seront ceux prédits par le système. Une comparaison des prédictions avec les annotations réelles de la taxonomie cible sera également présentée.

1. <https://github.com/deezer/muzeeglot>

Lors de la démonstration de Muzeeglot, nous nous attarderons plus particulièrement sur le sous-problème de *l'annotation multilingue d'entités musicales*. Nous présenterons un jeu de données de plus de 63 000 entités musicales (titres, artistes, albums...), collectées depuis DBpedia et annotées de genres musicaux de six langues différentes : Anglais, Espagnol, Français, Japonais, Néerlandais et Tchèque. Nous mettrons en avant l'efficacité de notre système à inférer les genres d'une langue à partir d'autres. Notons qu'il ne s'agit pas simplement d'un problème de traduction littérale des genres. En effet, les annotations intègrent les interprétations de rédacteurs de cultures et de pays différents ; par ailleurs, certains genres ne sont tout simplement pas traduits (*heavy metal* ne devient pas *métal lourd* en Français). La Figure 2 présente par exemple les prédictions de genres musicaux en Français pour le groupe Pink Floyd, à partir de ses annotations en Anglais et en Japonais au sein de DBpedia.

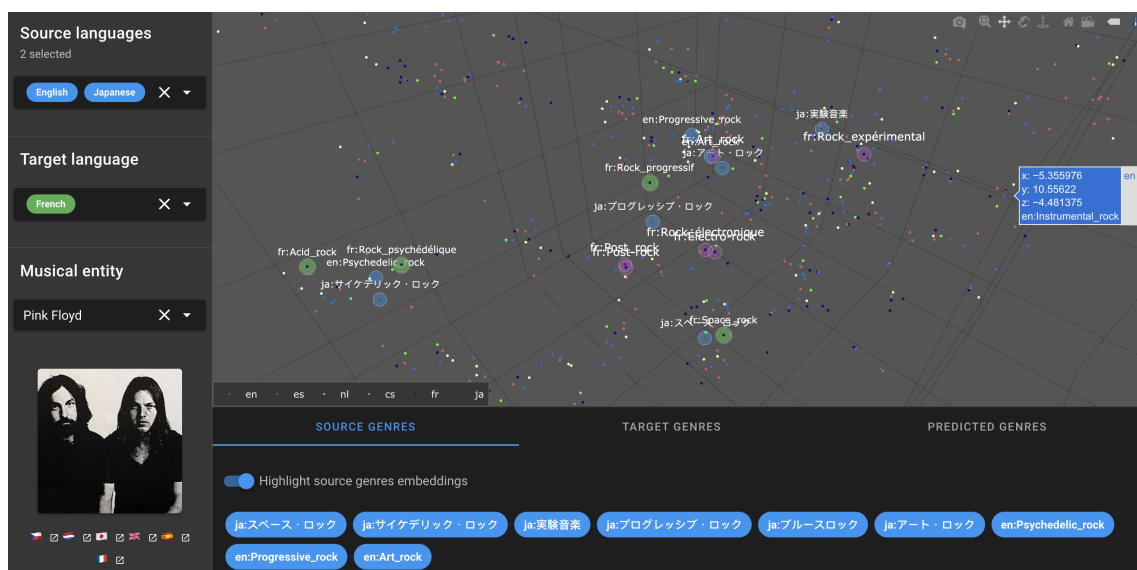


FIGURE 1 – Interface de Muzeeglot. Visualisation des genres musicaux de Pink Floyd, selon deux taxonomies *sources* correspondant aux annotations en Anglais et en Japonais au sein de DBpedia.

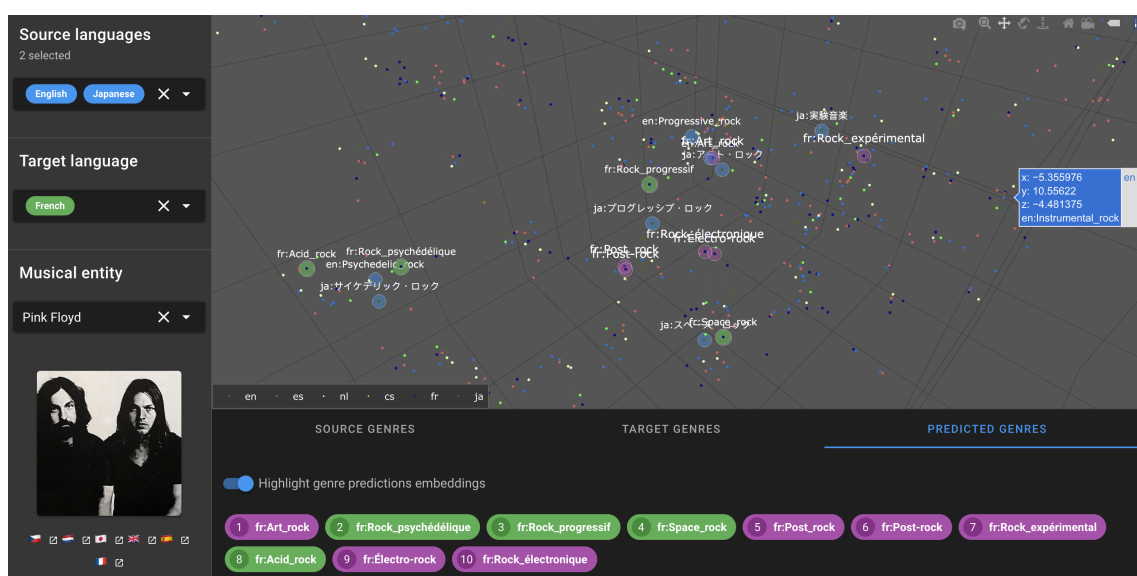


FIGURE 2 – Prédiction des genres musicaux de Pink Floyd en Français. Les prédictions représentées en vert coïncident avec les annotations de genres de Pink Floyd en Français au sein de DBpedia.

Références

- AUER S., BIZER C., KOBILAROV G., LEHMANN J., CYGANIAK R. & IVES Z. (2007). Dbpedia : A nucleus for a web of open data. In *The Semantic Web*. DOI : [10.1007/978-3-540-76298-0_52](https://doi.org/10.1007/978-3-540-76298-0_52).
- EPURE E. V., KHLIF A. & HENNEQUIN R. (2019). Leveraging knowledge bases and parallel annotations for music genre translation. In *Conference of the International Society of Music Information Retrieval*, ISMIR 2019. <http://archives.ismir.net/ismir2019/paper/000103.pdf>.
- FARUQUI M., DODGE J., JAUHAR S. K., DYER C., HOVY E. & SMITH N. A. (2015). Retrofitting word vectors to semantic lexicons. In *Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, NAACL-HLT 2015. DOI : [10.3115/v1/N15-1184](https://doi.org/10.3115/v1/N15-1184).
- GRAVE E., BOJANOWSKI P., GUPTA P., JOULIN A. & MIKOLOV T. (2018). Learning word vectors for 157 languages. In *International Conference on Language Resources and Evaluation*, LREC 2018. <https://www.aclweb.org/anthology/L18-1550>.
- HENNEQUIN R., ROYO-LETELIER J. & MOUSSALLAM M. (2018). Audio based disambiguation of music genre tags. In *Conference of the International Society of Music Information Retrieval*, ISMIR 2018. <http://archives.ismir.net/ismir2018/paper/000163.pdf>.
- JOULIN A., BOJANOWSKI P., MIKOLOV T., JÉGOU H. & GRAVE E. (2018). Loss in translation : Learning bilingual word mapping with a retrieval criterion. In *Conference on Empirical Methods in Natural Language Processing*, EMNLP 2018. DOI : [10.18653/v1/D18-1330](https://doi.org/10.18653/v1/D18-1330).
- LEE J. & DOWNIE J. (2013). K-pop genres : A cross-cultural exploration. In *Conference of the International Society on Music Information Retrieval*, ISMIR 2013. <http://archives.ismir.net/ismir2013/paper/000233.pdf>.
- MANDEL M., ECK D. & BENGIO Y. (2010). Learning tags that vary within a song. In *Conference of the International Society for Music Information Retrieval Conference*, ISMIR 2010. <http://archives.ismir.net/ismir2010/paper/000068.pdf>.
- MIKOLOV T., GRAVE E., BOJANOWSKI P., PUHRSCHE C. & JOULIN A. (2018). Advances in pre-training distributed word representations. In *International Conference on Language Resources and Evaluation*, LREC 2018. <https://www.aclweb.org/anthology/L18-1008>.
- PENNINGTON J., SOCHER R. & MANNING C. (2014). Glove : Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, EMNLP 2014. DOI : [10.3115/v1/D14-1162](https://doi.org/10.3115/v1/D14-1162).
- SCHEDL M. & FERWERDA B. (2017). Large-scale analysis of group-specific music genre taste from collaborative tags. In *IEEE International Symposium on Multimedia*, ISM 2017. DOI : [10.1109/ISM.2017.95](https://doi.org/10.1109/ISM.2017.95).
- SORDO M., CELMA O., BLECH M. & GUAUS E. (2008). The Quest for Musical Genres : Do the Experts and the Wisdom of Crowds Agree? In *Conference of the International Society on Music Information Retrieval*, ISMIR 2008. <http://archives.ismir.net/ismir2008/paper/000267.pdf>.

FlexEval, création de sites web légers pour des campagnes de tests perceptifs multimédias

Cédric Fayet¹ Alexis Blond¹ Grégoire Coulombel¹ Claude Simon¹
Damien Lolive¹ Gwénolé Lecorvé¹ Jonathan Chevelu¹ Sébastien Le Maguer²

(1) Univ Rennes, CNRS, IRISA, Lannion, France

(2) ADAPT Centre, Sigmedia Lab, EE Engineering, Trinity College Dublin, Dublin, Ireland

claudesimon.1@univ-rennes1.fr, {damien.lolive, gwenole.lecorve,
jonathan.chevelu}@irisa.fr, lemagues@tcd.ie

RÉSUMÉ

Nous présentons FlexEval, un outil de conception et déploiement de tests perceptifs multimédias sous la forme d'un site web léger. S'appuyant sur des technologies standards et ouvertes du web, notamment le framework Flask, FlexEval offre une grande souplesse de conception, des gages de pérennité, ainsi que le support de communautés actives d'utilisateurs. L'application est disponible en open-source via le dépôt Git <https://gitlab.inria.fr/expression/tools/flexeval>.

ABSTRACT

FlexEval, creation of light websites for multimedia perceptual test campaigns.

This paper presents FlexEval, an application which allows to design and deploy multimedia perceptual tests in the form of a light website. Using standard and open web technologies, especially the Flask framework, FlexEval offers great design flexibility, guarantees of sustainability, as well as support of active user communities. The application is open-source and available via the Git repository <https://gitlab.inria.fr/expression/tools/flexeval>.

MOTS-CLÉS : Tests perceptifs, multimedia, web.

KEYWORDS: Perceptual tests, multimedia, web.

1 Introduction

De nombreux domaines de recherche requièrent la récolte d'avis d'utilisateurs, notamment en traitement automatique des langues et de la parole. Par exemple, il peut s'agir pour les utilisateurs de se prononcer sur le naturel de signaux sonores produits par des systèmes de synthèse de parole, la validité des réponses d'un *chatbot*, l'expressivité d'avatars signant en langue des signes ou encore la pertinence de résultats d'un moteur de recherche. Cet article présente FlexEval, un outil permettant de créer et déployer de tels tests sans requérir de développements informatiques lourds ou de complexes étapes de configurations de la part du concepteur. Nous en donnons tout d'abord une vue d'ensemble (section 2), puis présentons un exemple simple d'un test perceptif (section 3). Nous détaillons ensuite des aspects liés à la flexibilité de FlexEval (section 4), puis les aspects juridiques (section 5).

2 Vue d'ensemble

FlexEval modélise un test perceptif de manière générique comme l'enchaînement de différentes scènes (*stages*) que le concepteur décide d'agencer selon son besoin. Les scènes disponibles sont

listées ci-dessous.

- **(Test)** La principale scène consiste, pour un utilisateur donné, en une succession d'*étapes* identiques lui présentant un ou plusieurs *échantillons* de données (textes, signaux sonores, vidéos. . .), une ou plusieurs questions et sollicitant ses réponses. Des modèles pour plusieurs grands types de base sont déjà définis (AB, MOS, MUSHRA. . .) et facilement adaptables.
- **(Authentification)** Un autre type de scène permet d'associer un identifiant à chaque utilisateur et de contrôler l'accès à une campagne en cours. Plusieurs variantes sont proposées : identification anonyme, identifiant spécifié par l'utilisateur, accès par invitation. . .
- **(Formulaire)** Il peut également être nécessaire d'adresser un questionnaire pour récolter des informations indépendantes de tout échantillon (par exemple, l'âge et la localité de l'utilisateur, ses commentaires en fin de test, sa connaissance *a priori* du domaine. . .).
- **(Page)** Enfin, il est également souvent utile de pouvoir passer par une scène explicative, ne nécessitant aucune action de l'utilisateur, typiquement un tutoriel. Ceci se fait sous la forme d'une simple page HTML.

Pour le concepteur, la mise en œuvre d'un test perceptif s'appuie sur la définition de l'agencement des scènes et de leurs éventuels paramètres, ainsi que sur la spécification de l'ensemble de données parmi lesquelles les échantillons seront tirés pendant le test. De son côté, l'application assure la gestion des identifiants, le brassage équilibré des données et la possibilité pour les utilisateurs de suspendre, puis reprendre le test comme ils veulent. Il en résulte un site web que le concepteur peut placer de manière autonome sur un serveur de son choix, qui, à l'issue de la campagne de test, produira un fichier de tous les résultats obtenus, sous la forme d'un fichier CSV ou d'une base de données SQLite. L'exportation de média est possible sous la forme de champ BLOB ¹ au format binaire.

FlexEval est indépendant de logiciels tiers et facile à installer car il s'appuie sur des logiciels standard de développement web. Du côté serveur, l'application utilise le framework Flask ², fondé sur Python, et le gestionnaire de base de données embarquées SQLite. Du côté client, les technologies utilisées sont HTML5, CSS et JavaScript (notamment JQuery). L'ensemble du code est récupérable à l'adresse <https://gitlab.inria.fr/expression/tools/flexeval>

3 Exemple de test perceptif

Prenons l'exemple d'un test dans lequel on souhaite comparer des phrases produites par 2 systèmes de synthèse de la parole, référencés comme les systèmes *hyp* et *ref*. Pour un testeur donné, nous souhaitons que la session de test débute par une authentification, suivie d'un questionnaire, puis 40 étapes d'un test de type MOS (*Mean Opinion Score*). À chaque étape, nous supposons vouloir présenter un texte et un échantillon de parole correspondant, puis obtenir deux notes entre 1 et 5 de la part du testeur, par exemple sur les degrés de spontanéité et d'intelligibilité du signal de parole. Pour un ensemble de textes donnés, il existe un échantillon de parole associé pour chaque système de synthèse testé. Les échantillons présentés au testeur viendront aléatoirement soit de l'un, soit de l'autre. Une fois les 40 étapes terminées, le test se conclut par une page de fin.

Pour lancer le site d'une campagne d'évaluation, FlexEval s'appuie sur un répertoire créé par le concepteur et qui contient toutes les informations utiles. L'arborescence de ce répertoire est illustré par la figure 1. Nous supposons que les données associées à chaque système sont listées au sein de fichiers CSV. Par exemple, nous créons deux fichiers *hyp.csv* et *ref.csv* qui, pour chaque ligne, liste un texte et le chemin vers le fichier WAV associé. Sur cette base, la figure 2 montre les deux fichiers de configuration qui permettent de spécifier à FlexEval la campagne d'évaluation. D'une part, la définition de l'enchaînement des scènes prend la forme du fichier *structure.json*. Les scènes

1. <https://developer.mozilla.org/fr/docs/Web/API/Blob>

2. <https://flask.palletsprojects.com/>

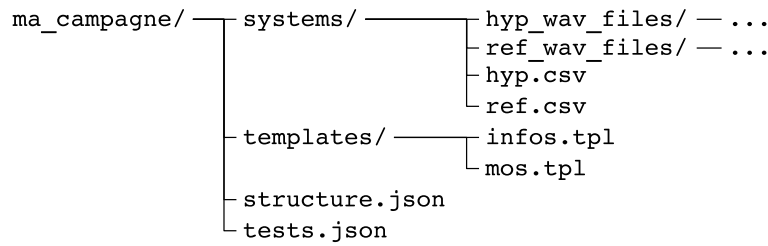


FIGURE 1 – Arborescence de fichiers de la campagne d'évaluation.

<pre> 1 { "stages": { 2 "debut":{"type": "auth_by_invite", "next":"infos", ...}, 3 "fin": {"type": "page:user", ...}, 4 "infos": {"type": "form", "template": "...", "next": "mos", ...}, 5 "mos": {"type": "test", "next": "fin", 6 "template": "mos.tpl", "nb_steps": 40, ...} 7 }, 8 "entrypoint": "debut" }</pre>	<pre> 1 { 2 "mos": 3 [4 {"name": "hyp", "data": "hyp.csv"}, 5 {"name": "ref", "data": "ref.csv"} 6] 7 }</pre>
structure.json	tests.json

FIGURE 2 – Configuration du scénario et des données du test perceptif.

sont identifiées par un nom et leur ordre de succession est indiqué par le champ `next`. Un squelette de page web peut être associée à chaque scène lorsque nécessaire. Par exemple, pour la phase `mos`, nous souhaitons utiliser un patron de fichier HTML (*template*) que nous avons défini : `mos.tpl`. Un exemple de rendu de ce patron est donné par la figure 3. Sans précision, un patron fourni par défaut est utilisé. D'autre part, la configuration pour l'accès aux données (fichiers CSV) de chaque système est définie dans `tests.json`. Les noms associés à chaque système seront ceux utilisés dans la base de données et au moment de l'export des résultats au format CSV. Ici, nous référençons nos systèmes *hyp* et *ref* et les associations à leur fichier CSV respectif.

Les différents fichiers mis en place, le concepteur lance le serveur web de FlexEval en indiquant le chemin vers le répertoire `ma_campagne/`, éventuellement en indiquant l'IP et le port auquel le site web sera accessible : `python3 run.py -ip 123.45.67.89 -port 8080 /chemin/vers/ma_campagne`. Le concepteur communique alors l'URL du site de la campagne. Une fois la campagne terminée, le concepteur récupère les réponses collectées via une interface d'administration dédiée ou directement à la racine du répertoire `macampagne/`. Le serveur web peut enfin être interrompu.

4 Éléments de flexibilité

FlexEval autorise une grande souplesse dans la création d'un test. Au niveau de la conception globale du test, l'emploi d'un fichier de configuration `structure.json` permet de décrire de manière flexible et intuitive un test sous la forme d'un scénario. Il n'y a pas de contrainte sur l'enchaînement des scènes. Si aucune phase d'authentification n'est prévue, l'outil en prévoira une par défaut, transparente pour l'utilisateur. Il est également possible de spécifier plusieurs scènes de tests successives.

Ensuite, le recours à des patrons de fichiers HTML et de possibles bibliothèques JavaScript permet une grande liberté d'interaction avec l'utilisateur que ce soit en terme de présentation des contenus (textes, vidéos, images...) que de récolte des réponses de l'utilisateur (multiples composants de formulaire, captation audio, vidéo...).

En outre, FlexEval permet l'inclusion de segments de code en Python au sein des patrons de fichiers

Step 7/40

Question: Between 1 and 5, how do you judge the pronunciation of the following sample regarding:

- **spontaneity**,
- and **intelligibility**?

Remark: All the signals are buzzy due to the synthesis technology used, this is independent of the experiments.

Text "well definitely this happened to me when I was in the"

Sample ▶ ● 0:00 / 0:04 ◀

Spontaneity

1: Neutral/formal

5: Perfectly spontaneous

3

Intelligibility

1: Unintelligible

5: Perfectly intelligible

3

Next

FIGURE 3 – Copie d’écran d’une étape d’un test MOS.

HTML, selon le langage de squelette Jinja³. Ces segments ont accès à des variables définies par FlexEval concernant l’utilisateur courant, la scène en cours (par exemple, pour une scène de test, le nombre d’étapes, le nombre de systèmes ou les informations sur un échantillon) ou encore des aspects explicitement fournis par le concepteur à travers le fichier `structure.json` (par exemple, le titre de la campagne, l’auteur des expériences...). Par ailleurs, les patrons peuvent inclure d’autres patrons, ce qui permet une meilleure réutilisation. Par exemple, il est possible d’inclure un pied de page dont le contenu (par exemple, des logos) est défini dans un autre fichier.

Enfin, il est aisé de créer ses propres types de scènes en Python en s’inspirant de celles existantes ou de définir des spécialisations de certains types par un mécanisme d’héritage, sans nécessité de modifier des fichiers existants de FlexEval.

5 Aspects juridiques

Tout responsable de l’utilisation de FlexEval au sein de son organisation doit se conformer aux obligations du Règlement Général sur la Protection des Données (RGPD⁴). Par exemple, au début du test, une page spécifiques doit présenter les principaux éléments liés à l’emploi de FlexEval mis en conformité avec le RGPD. Le consentement de l’utilisateur est rendu obligatoire à l’aide d’une case à cocher. Nous avons automatisé la génération du patron HTML de cette page spécifique en l’associant au document `legal.json`. Ce document est à compléter avec les informations exigées par le RGPD comme par exemple le nom et les coordonnées de l’organisation, les objectifs, la finalité et la nature des données collectées, les mesures de confidentialité, de droit d’accès, de sécurité ou de durée de conservation de ces données. On peut aussi y associer des données de clauses de CGU⁵).

3. <http://jinja.pocoo.org/>

4. <https://www.economie.gouv.fr/particuliers/reglement-general-protection-des-donnees-rgpd>

5. Conditions Générales d’Utilisation

Vers une analyse automatique de la perception relative à un lieu

Hélène Flamein¹, Iris Eshkol-Taravella²,

(1) Laboratoire Ligérien de Linguistique (CNRS UMR 7270), Université d'Orléans

(2) MoDyco (CNRS UMR 7114), Université Paris Nanterre

helene.flamein@univ-orleans.fr, ieshkolt@parisnanterre.fr

RESUME

Le travail présenté s'intéresse à la perception qu'ont les habitants de leur ville en se fondant sur un corpus de conversations orales spontanées. La chaîne de traitement conditionnant l'analyse de la perception se décompose en trois étapes : la détection des noms de lieux, l'analyse de la perception identifiée et la visualisation cartographique des informations extraites.

ABSTRACT

Towards an Automatic Analysis of Place Perception.

The work presented deals with the perception that the inhabitants have of their city based on a corpus of spontaneous oral conversations. The processing chain that conditions the analysis of perception is divided into three steps: the detection of place names, the analysis of the identified perception and the cartographic visualization of the extracted information.

MOTS-CLES : Lieux, Subjectivité, Polarité, Oral transcrit, ESLO.

KEYWORDS: Place, Subjectivity, Polarity, Oral transcript, ESLO.

1 Introduction

Le travail présenté s'intéresse à la perception qu'ont les habitants de leur ville en se fondant sur un corpus de conversations orales spontanées tiré d'ESLO (Enquête SocioLinguistique à Orléans)¹. La perception est donc analysée à travers trois étapes : la détection des noms de lieux mentionnés dans le corpus, l'analyse de la perception dans le contexte des lieux identifiés et les visualisations des informations extraites tout au long du traitement.

2 Détection des lieux dans l'oral transcrit

Le processus de détection automatique de noms de lieux dans le corpus oral transcrit est décrit dans Flamein & Eshkol-Taravella (2020). D'une manière générale, le système s'appuie sur l'exploitation de ressources lexicales Geonames, GEOFLA et Data.gouv.fr qui référencient des noms de lieux normalisés avec leurs coordonnées géographiques. En l'état, ces ressources ne suffisent pas pour l'annotation exhaustive des lieux puisqu'elles ne recensent que les noms officiels, conventionnels des lieux. Pour détecter les noms de lieu non normalisés, le système génère des variantes de noms de

¹

<http://eslo.huma-num.fr/>

lieux afin d'enrichir les ressources lexicales. Les règles utilisées pour cela sont établies à partir de l'observation des pratiques des locuteurs du corpus en ce qui concerne la dénomination des lieux. Par exemple, dans les noms de voie, les locuteurs ne conservent que le dernier mot (rue Gauguin au lieu de la rue Paul Gauguin) ou groupe prépositionnel (rue de Sonis au lieu de la rue du Général de Sonis) composant le nom officiel ainsi que le mot caractérisant le type de voie. Dans le cas des noms de villes composés de trois mots ou plus, c'est plutôt le premier terme qui est conservé et les derniers termes qui sont supprimés (La Ferté Saint-Aubin pour La Ferté). De nouvelles entrées sont donc générées et ajoutées au lexique sur la base de ces observations. Le lexique ainsi enrichi est ainsi appliqué sur le corpus et associé à des patrons permettant l'identification des noms de lieux tronqués (*rue de la Rép-* pour la *rue de République*), ou même inventés (*rue de la gare, boulangerie du coin*, etc.).

Le module développé est évalué sur un corpus de référence composé de 15 transcriptions, référençant 2292 noms de lieux, parmi lesquels 549 (24%) varient par rapport à la norme. Le module obtient un Rappel de 0,90, une Précision de 0,93 et une F-Mesure de 0,91. Cette évaluation porte sur les frontières du nom du lieu et montre des performances satisfaisantes pour identifier des lieux sous leurs formes officielles et leurs variantes. L'annotation des lieux sert d'ancrage pour la deuxième étape d'analyse de la perception effectuée par apprentissage automatique.

3 Traitement de la subjectivité et de la polarité

La perception est une notion vaste que la linguistique, la psychologie et même les sciences de l'information ont tenté de circonscrire. Du point de vue du TAL, la perception est abordée sous l'angle de l'analyse de sentiment et de la fouille d'opinion (Pak & Paroubek, 2010 ; Marchand, 2015 ; Karaoui et al., 2019). Ces travaux se fondent principalement sur les données issues du Web. Les données orales sont moins facilement disponibles mais présentent tout autant d'enjeux pour la problématique de la fouille d'opinion et de l'analyse de sentiments.

L'analyse de la perception relative à Orléans s'appuie sur la première phase de détection des lieux dans les transcriptions du corpus ESLO et est réalisée grâce à des techniques d'apprentissage automatique supervisé. La classification des énoncés est réalisée par apprentissage automatique sur un corpus de référence annoté manuellement² et divisé en trois parties : corpus d'entraînement, de test et d'évaluation. Afin d'entraîner un modèle pour la détection de la subjectivité et de la polarité, différents *features* sont envisagés : la lemmatisation et l'étiquetage morphosyntaxique ainsi que le calcul d'un score de polarité et d'un score d'émotion³. Le TF-IDF et Word2Vec sont des méthodes possibles pour créer une représentation vectorielle des segments à analyser. Enfin deux classifieurs, RandomForest et SVM, ont été retenus pour réaliser la classification attendue.

Les expériences mêlant les différents *features*, les méthodes de représentations vectorielles et les classifieurs ont été menées pour sélectionner la combinaison la plus performante. L'utilisation du classifieur SVM, la représentation vectorielle des segments lemmatisés avec le TF-IDF et l'utilisation de tous les features disponibles constituent le modèle le plus efficace. Les expériences révèlent que le même modèle est le plus efficace pour distinguer les énoncés subjectifs des énoncés objectifs mais aussi pour détecter la polarité des énoncés. Pour confirmer les performances de ce modèle, celui-ci est évalué sur le corpus de test. Les résultats de cette évaluation sont présentés dans

² Constitué de trente transcriptions du corpus ESLO, annotés en lieux (4519 mentions)

³ Ces scores sont définis à partir de l'exploitation du lexique FEEL (Abdaoui et al. 2017)

le tableau 1. Pour la tâche de détection de la subjectivité, le modèle obtient une macro-average de 0,77 et une macro-average de 0,76 pour le traitement de la polarité.

Classifieur	Features	Cible	Macro average	Précision	Rappel	F-mesure
TF-IDF + SVM	Score polarité + Score émotions + POS	subj.	0,77	0,69	0,56	0,63
		obj.		0,77	0,89	0,83
		pos.	0,76	0,78	0,91	0,82
		nég.		0,67	0,46	0,54

TABLEAU 1 : Modèle retenu pour la détection de la subjectivité et de la polarité

4 Visualisation de la perception

L'objectif de la dernière étape de notre travail est de représenter visuellement les résultats des analyses réalisées jusqu'ici. Il s'agit de mettre en valeur les informations extraites de notre corpus afin de faciliter leur manipulation et les rendre accessibles. Dans une autre mesure, la visualisation contribue à faire émerger les relations qui unissent les différentes informations détectées. La visualisation de la perception par les Orléanais de leur ville se concrétise donc avec la mise en place d'un Système d'Information Géographique (SIG)⁴ et la création de cartes. Pour cela, nous utilisons l'outil ArcGIS Online⁵, développé par Esri et disponible en ligne. Plusieurs couches d'informations sont projetées dans le système et permettent de figurer l'ensemble des lieux détectés en fonction des métadonnées qui leurs sont associées au cours du traitement et des déclarations qui leurs sont relatives. La figure 1 présente les lieux situés dans le centre-ville d'Orléans en fonction de la polarité. Un ratio correspondant à la part de déclarations positives et négatives est calculé et corrélé avec le nombre total de déclarations subjectives exprimées à propos du lieu. Plus le nombre de déclarations est important, plus le ratio de la polarité est significatif.

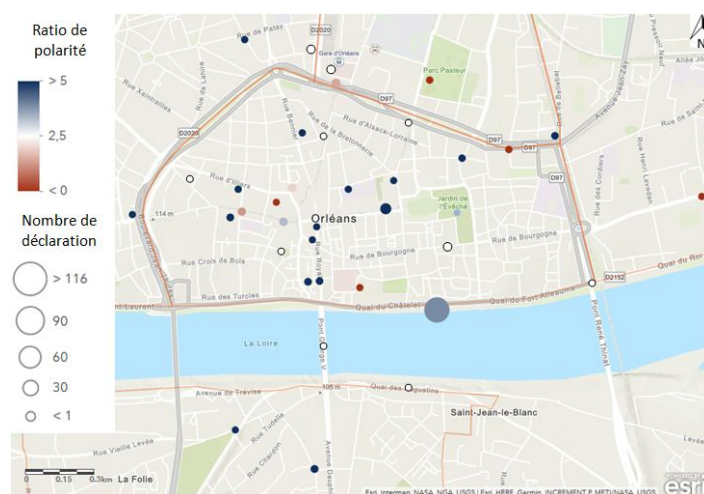


FIGURE 1 : Répartition de la polarité à propos des lieux du centre-ville d'Orléans

⁴ SIG pour la visualisation de la perception de la ville d'Orléans – <http://arcg.is/1v9DaC>
⁵ <https://www.arcgis.com/index.html>

5 Conclusion et perspectives

Nous avons présenté une méthodologie pour la détection de la perception qu'ont les habitants de leur ville. Afin de procéder à l'analyse de cette perception, nous avons d'abord développé un système de détection des noms de lieux mentionnés tout en tenant compte des spécificités du corpus. A partir des détections réalisées, un modèle pour la détection de la subjectivité et la détermination de la polarité a été entraîné. Finalement, l'ensemble des informations identifiées tout au long du traitement ont été projetées dans un SIG afin de donner corps à l'image que se font les orléanais de leur ville.

Références

ABDAOUI A., AZE J., BRINGAY S. & PONCELET P. (2017). FEEL : A French Expanded Emotion Lexicon. *Language Resources and Evaluation*, 51(3), pp. 833–855. DOI : <https://doi.org/10.1007/s10579-016-9364-5>

FLAMEIN H. & ESHKOL-TARAVELLA I. (2020). Noms de lieux dans le corpus de français parlé : Une approche symbolique pour un traitement automatisé. *Le français moderne 2020*, n.1

KARAOUI J. & BENAMARA F. & Véronique Moriceau. (2019) *Détection automatique de l'ironie: Application à la fouille d'opinion dans les microblogs et les médias sociaux*. ISTE Group.

MARCHAND M. (2015). *Domaines et fouille d'opinion: une étude des marqueurs multipolaires au niveau du texte*. Thèse de doctorat, Université Paris Sud - Paris XI. [tel-01157951](tel:01157951)

NOVAKOVA I. (2019). *Le lexique des émotions*. UGA Éditions.

PAK A. & PAROUBEK P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *LREc*, 10, pp. 1320–1326. <https://www.aclweb.org/anthology/L10-1263/>

ZHANG L. (2012). *Analyse automatique d'opinion : Problématique de l'intensité et de la négation pour l'application à un corpus journalistique*. Thèse de doctorat, Université de Caen. [tel-00777603](tel:00777603)

Un prototype en ligne pour la prédiction du niveau de compétence en anglais des productions écrites

Thomas Gaillat¹ Nicolas Ballier² Annanda Sousa³ Manon Bouyé²
Andrew Simpkin³ Bernardo Stearns³ Manel Zarrouk⁴

(1) LIDILE, Université de Rennes 2, 35000 Rennes, France

(2) Insight Centre for Data analytics, NUI Galway, Irlande

(3) CLILLAC-ARP, Université de Paris, 75013 Paris, France

(4) LIPN, Université Sorbonne Paris Nord, 93430, France

thomas.gaillat@univ-rennes2.fr, nicolas.ballier@u-paris.fr,
a.defreitassousa@nuigalway.ie, mbouye@eila.univ-paris-diderot.fr,
andrew.simpkin@nuigalway.ie, zarrouk@lipn.univ-paris13.fr,
bernardo.stearns@nuigalway.ie

RÉSUMÉ

Cet article décrit un prototype axé sur la prédiction du niveau de compétence des apprenants de l'anglais. Le système repose sur un modèle d'apprentissage supervisé, couplé à une interface web.

ABSTRACT

A prototype for web-based prediction of English proficiency levels in writings.

This paper describes a proof-of-concept system focused on proficiency level prediction in learners of English. The system relies on a supervised learning model coupled with a web interface for users.

MOTS-CLÉS : CECRL, Système d'évaluation automatique, anglais d'apprenant, complexité.

KEYWORDS: CEFR, Automatic Essay Scoring, learner English, linguistic complexity.

1 Introduction

Le système présenté¹ était destiné à servir de preuve de concept pour l'évaluation automatique du niveau de langue en anglais. Nos recherches visent à identifier les caractéristiques linguistiques et à les intégrer dans un système fondé sur l'intelligence artificielle (IA). L'objectif est de créer un système permettant d'analyser les essais des apprenants de l'anglais et de les mettre en correspondance avec des niveaux spécifiques des niveaux de langue du Cadre Européen Commun de Référence pour les langues (CECRL, [European Council \(2001\)](#)). Nous présentons un système fonctionnant par apprentissage supervisé. Le modèle s'appuie sur les caractéristiques des textes, indépendamment des erreurs, pour en construire une représentation multidimensionnelle sous la forme de traits critériés ([Hawkins & Filipović, 2012](#)). Les caractéristiques retenues comprennent la lisibilité et les mesures de complexité utilisées dans le domaine de l'exploration de textes et du TAL.

1. Cette démonstration synthétise en français des publications précédentes. Les détails concernant l'approche linguistique, la modélisation, les résultats en terme de précision et l'architecture ont été publiés dans ([Sousa et al., 2020](#)). La chaîne de traitement est en python (3.6) et ses parties redistribuables le sont avec la licence Creative Commons.

2 Description du système

Le système a été entraîné sur une base de données² de plus de 40 000 textes (environ 3 298 343 tokens), qui ont déjà été étiquetés et annotés en niveaux (Geertzen *et al.*, 2013). A partir des 769 caractéristiques des différentes dimensions linguistiques, le meilleur modèle s’est avéré être une régression logistique. Le modèle repose sur les caractéristiques des productions, indépendamment des erreurs commises par les apprenants (par opposition à une analyse fondée sur les erreurs annotées, cf. Ballier *et al.* (2019)). Plusieurs outils (cf. Tableau 1) analysant les niveaux de complexité linguistique sont concaténés dans une chaîne de traitement pour construire une représentation multidimensionnelle des caractéristiques des essais écrits. L’évaluation du système a montré une précision de 82% (cf. Gaillat *et al.* (submitted)).

Les productions des apprenants sont souvent évaluées en fonction des trois dimensions que sont la complexité, la précision et la fluidité ("fluency") (Housen *et al.*, 2012). Les métriques proposées opérationnalisent la dimension de la complexité linguistique et celle de la précision des écrits. Les métriques de complexité portent sur les dimensions syntaxiques, lexicales, discursives et psycholinguistiques sous la forme d’indices de lisibilité et d’îlots de fiabilité ("reliability islands" unités phraséologiques récurrentes, Dechert, Hans-Wilhelm (1983)). Les métriques de complexité syntaxique traditionnelles abordent généralement les textes de manière syntagmatique. Nous proposons de nouvelles métriques fondées sur les rapports paradigmatiques qu’entretiennent des formes linguistiques entre elles. Nous formalisons des microsystèmes linguistiques sous la forme de ratios d’usage de ces formes les unes par rapport aux autres et indépendamment des autres formes des textes. Le principe est de saisir les variations d’usage des formes par les apprenants (Gaillat *et al.*, submitted). L’outil L2SCA a été modifié en ce sens et est disponible à l’adresse du projet.

Outils	Dimension	Exemples de métriques
L2SCA modified (Lu, 2014)	Complexité syntaxique	Mean Length Sentence (MLS), Microsystèmes
LCA (Lu, 2014)	Diversité lexicale	Type Token Ratio (TTR)
TAALES (Kyle <i>et al.</i> , 2018)	Diversité et Sophistication lexicale	Fréquences de mots et écarts types dans un corpus
TAASC (Kyle <i>et al.</i> , 2018)	Sophistication syntaxique	Nombre de prépositions par groupe nominal
TAACO (Kyle <i>et al.</i> , 2018)	Complexité discursive : Cohésion textuelle	Chevauchements lexicaux, répétitions de pronoms
Pyenchant 2.0.0 (Kelly, 2016)	Erreurs d’orthographe : dictionnaire Aspell	Fréquences d’erreurs
Textstat (Bansal, 2018)	Lisibilité, Phraséologie	Indices de type Dale_Chall, fréquences de N-grams

TABLE 1 – Récapitulatif des principaux outils de la chaîne de traitement

Ce prototype est composé de deux modules. Le premier est une interface web utilisateur³ en accès libre permettant la prédiction de niveau CECRL (Cf. Figure 1). Les apprenants saisissent leurs textes avant de recevoir un feedback sur la classe CECRL estimée et sa probabilité. Cette infrastructure est

2. Les données tirées du corpus EFCAMDAT ont été annotées indépendamment des équipes de recherche de Cambridge et d’EF Education.

3. Cf. l’URL du projet : www.clillac-arp.univ-paris-diderot.fr/projets/ulyssse2019.

composée de modules Docker (Merkel, 2014) interconnectés permettant la production des métriques et la classification par le modèle.

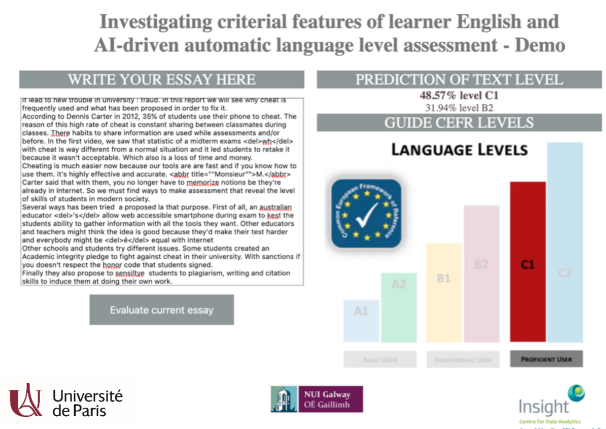


FIGURE 1 – L’interface utilisateur pour les apprenants de l’anglais

Le second module est une chaîne de traitement en ligne de commande. Au contraire de l’interface utilisateur, ce module permet de traiter des fichiers textes par lots et ainsi de créer des jeux de données représentant les valeurs des métriques pour les textes. Du fait de la mise en oeuvre d’un outil dépendant d’un corpus propriétaire, seule une version plus restreinte de la chaîne de traitement, telle que décrite dans (Sousa *et al.*, 2020) est disponible pour la communauté.

3 Développements et perspectives

Nous envisageons deux voies d’amélioration du prototype. Une première extension concerne l’ajout du traitement de la parole en ayant recours à des outils existants de retranscription automatique tels que la librairie Python SpeechRecognition (Zhang, 2017). La seconde extension concernera la variété des métriques. On explorera d’autres métriques de mesure des microsystemes linguistiques, notamment pour l’analyse de la morphologie (Brezina & Pallotti, 2019), afin de mieux analyser les variations d’usages propres aux apprenants. En outre, les métriques seront exploitées dans le cadre de visualisations permettant de comparer des individus avec des profils types d’apprenants. Cela nous mènera à étudier l’impact de ce type d’outil en situation pédagogique d’enseignement. Les feedbacks sur les propriétés positives (Hawkins & Filipović, 2012) de la langue peuvent avoir des conséquences sur la précision dans l’expression (*accuracy*), ce qui est important à évaluer. Ce prototype est à considérer comme un moyen d’inventer de nouveaux outils pour aider les enseignants dans leur pratique. Dans le cadre des diagnostics d’apprentissage, ils/elles bénéficieraient d’outils d’analyse faciles à utiliser, et qui objectivent les progrès de leurs apprenants.

Remerciements

Nous tenons à remercier l’Irish Council et les Ministères français des Affaires étrangères et de la Recherche pour leur soutien dans le cadre du programme de financement Ulysse PHC 2019 (projet 43121RJ). Nous sommes reconnaissants à Kristopher Kyle et Scott A. Crossley de nous avoir fourni le code source des outils TAALES, TAACO et TAASC. Nos remerciements vont également à Xiaofei Lu pour avoir fourni L2SCA et LSA.

Références

- BALLIER N., GAILLAT T., SIMPKIN A., STEARNS B., BOUYÉ M. & ZARROUK M. (2019). A supervised learning model for the automatic assessment of language levels based on learner errors. In *European Conference on Technology Enhanced Learning*, p. 308–320 : Springer.
- BANSAL S. (2018). Textstat Python package. Accessible sur <https://github.com/shivam5992/textstat>.
- BREZINA V. & PALLOTTI G. (2019). Morphological complexity in written l2 texts. *Second language research*, **35**(1), 99–119.
- DECHERT, HANS-WILHELM (1983). How a story is done in a second language. In C. FÆRCH & G. KASPER, Éd., *Strategies in interlanguage communication*, p. 175–195. London ; New York : Longman. OCLC : 644977107.
- EUROPEAN COUNCIL (2001). *Common European Framework of Reference for Languages : Learning, teaching, assessment*. Cambridge : Cambridge University Press.
- GAILLAT T., SIMPKIN A., BALLIER N., STEARNS B., SOUSA A., BOUYÉ M. & ZARROUK M. (submitted). Predicting CEFR levels in learners of English : the use of microsystem criterial features in a machine learning approach. *Journal With Anonymous Submission*.
- GEERTZEN J., ALEXOPOULOU T. & KORHONEN A. (2013). Automatic Linguistic Annotation of Large Scale L2 Databases : The EF-Cambridge Open Language Database (EFCamDat). In R. T. MILLER, K. I. MARTIN, C. M. EDDINGTON, A. HENERY, N. MIGUEL, A. TSENG, A. TUNINETTI & D. WALTER, Éd., *Proceedings of the 31st Second Language Research Forum*, Carnegie Mellon : Cascadilla Press.
- HAWKINS J. A. & FILIPOVIĆ L. (2012). *Criterial Features in L2 English : Specifying the Reference Levels of the Common European Framework*. United Kingdom : Cambridge University Press.
- HOUSEN A., KUIKEN F. & VEDDER I., Éd. (2012). *Dimensions of L2 performance and proficiency : complexity, accuracy and fluency in SLA*, volume 32 de *Language Learning & Language Teaching (LL<)*. Amsterdam, Pays-Bas, Etats-Unis d'Amérique : John Benjamins Publishing Company.
- KELLY R. (2016). PyEnchant a spellchecking library for Python. Accessible sur <https://pythonhosted.org/pyenchant>.
- KYLE K., CROSSLEY S. & BERGER C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES) : version 2.0. *Behavior Research Methods*, **50**(3), 1030–1046.
- LU X. (2014). *Computational Methods for Corpus Annotation and Analysis*. Dordrecht : Springer.
- MERKEL D. (2014). Docker : lightweight linux containers for consistent development and deployment. *Linux journal*, **2014**(239), 2.
- SOUSA A., BALLIER N., GAILLAT T., STEARNS B., ZARROUK M., SIMPKIN A. & BOUYÉ M. (2020). From Linguistic Research Projects to Language Technology Platforms : A Case Study in learner data. In *Proceedings of the 1st International Workshop on Language Technology Platforms IWLTP 2020*, Marseille : LREC / ACL.
- ZHANG A. (2017). Speech recognition (version 3.8) python library. Accessible sur <https://pypi.org/project/SpeechRecognition/>.

LEX : un jeu avec finalité d'acquisition de ressources lexicales

Asheesh Gulati

LATL, Université de Genève, Suisse

Asheesh.Gulati@etu.unige.ch

RÉSUMÉ

LEX est un *jeu avec un but* développé dans l'optique d'explorer plus avant les éléments et principes de la conception de jeux tels qu'ils sont pratiqués dans l'industrie vidéoludique, pour les mettre au service de la conception de jeux sérieux. Le premier prototype repose sur un mode *bac à sable* pour faire appel à la créativité du joueur et renforcer l'immersion ludique.

ABSTRACT

LEX : a game with the purpose of lexical resource acquisition

LEX is a *game with a purpose* developed with the intent of further exploring the elements and concepts of game design as they are applied to video games, in order to apply them to the design of serious games. The first prototype offers a *sandbox* mode meant to stimulate the player's creativity and reinforce the sense of immersion.

MOTS-CLÉS : jeu sérieux, jeu avec un but, sciences participatives, ressources lexicales.

KEYWORDS: serious game, game with a purpose, citizen science, lexical resources.

1 Le project

Le pitch Dans LEX, le joueur incarne un féru de technologie œuvrant à la réalisation d'un robot, prénommé A-LEX, capable de communiquer par écrit. A-LEX interagit avec le joueur en lui posant des questions, et chaque réponse contribue à améliorer son vocabulaire. Des défis ponctuels sont disponibles afin de mesurer les capacités d'A-LEX, et le joueur peut choisir de s'y confronter seul ou avec d'autres joueurs.

Conçu à l'origine comme un *jeu avec un but* (*game with a purpose*, von Ahn & Dabbish (2008)) inspiré de JEUXDEMOTS (Lafourcade & Joubert, 2013), le projet sous sa forme actuelle peut tout à fait être approché comme un jeu sérieux dans une perspective didactique (Abt, 1987), destiné en l'occurrence à l'étude des techniques du traitement automatique des langues naturelles.

Chaque session de jeu se déroule en mode *bac à sable*, permettant aux joueurs de librement explorer les possibilités offertes, et expérimenter à loisir les outils disponibles. Le prototype actuel n'en est qu'à ses débuts, mais propose déjà une base pour l'exploration de nouvelles pistes de recherche dans la conception de jeux sérieux.

2 Les caractéristiques

LEX repose sur les principes suivants, et les incorpore comme éléments de sa mécanique de jeu et de sa jouabilité.

Motivation intrinsèque Contrairement à la majorité des *jeux avec un but*, qui exploitent une motivation qualifiée d’extrinsèque, liée à des éléments externes à l’activité ludique principale tels que scores, classements, badges/distinctions, etc., la motivation intrinsèque est ici privilégiée, et les objectifs « sérieux » sont valorisés et influencent directement l’expérience ludique (McGonigal, 2011). Les effets des actions du joueur sont clairement perceptibles, ce qui contribue à l’immersion ludique. Il s’agit là d’un des composants menant à l’état de *flux* (Csikszentmihalyi, 2008), qui se caractérise par l’absorption totale d’une personne par son occupation, et qui est considéré comme un objectif désirable dans la conception de jeux (Salen & Zimmerman, 2003).

Curiosité et créativité Le mode de jeu par défaut est un *bac à sable*, dans lequel aucun objectif prédéfini n’est imposé au joueur. Celui-ci est ainsi invité à donner libre cours à sa curiosité et créativité. Ce mode de jeu est en parfaite adéquation avec les visées des *jeux avec un but*, et des jeux sérieux en général (Squire, 2008; Schrier, 2018), mais nécessite néanmoins d’être étudié plus avant d’un point de vue scientifique.

Compétition et coopération Le conflit est un élément fondamental de tout jeu, et tous les jeux sont de ce fait compétitifs (Salen & Zimmerman, 2003). Toutefois, cela ne signifie aucunement que la relation entre les joueurs doit forcément être de nature compétitive : de nombreux jeux impliquent une coopération entre joueurs, et les mettent en conflit avec le système lui-même. Ces deux types de relations sont propices à des expériences ludiques très différentes, et peuvent être exploitées efficacement dans le contexte des *jeux avec un but* pour concevoir des activités adaptées aux visées recherchées (Tuite, 2014).

3 Le prototype

À ce stade de son développement, LEX propose au joueur de converser librement avec A-LEX, ainsi que de l’aider à améliorer son vocabulaire au travers de questions spécifiques. Les fonctionnalités en cours de mise en œuvre sont décrites dans la section suivante.

Chaque joueur se voit attribuer « son » robot. Celui-ci est initialisé à partir d’un modèle de référence (*baseline*), puis apprend au fur et à mesure de ses interactions avec « son » joueur.

Conversation libre La conversation libre avec A-LEX est destinée à renforcer l’immersion ludique, en permettant non seulement au joueur de créer un lien émotionnel avec le robot, mais aussi d’apprécier le résultat des efforts consacrés à lui enseigner de nouveaux termes.

Enseignement guidé À l'image de JEUXDEMOTS, l'activité principale de LEX consiste à proposer des termes répondant à des demandes précises d'A-LEX. La jouabilité est toutefois très différente :

- Chaque demande est formulée dans un registre familier et n'attend qu'une seule réponse.
- La durée des demandes n'est pas limitée, et le joueur peut prendre son temps pour proposer un terme, ou passer à la demande suivante sans avoir à fournir un terme.
- Aucun score mesurant la performance du joueur n'est calculé, et le seul retour consiste en une animation indiquant que le terme a bien été assimilé.

De temps en temps, la demande d'A-LEX n'attend pas un terme en réponse, mais une validation par oui ou non. Ce type de demandes est en rapport avec un processus de vérification des termes pour lesquels il existe une incertitude. Ces termes ne sont pas limités aux contributions préalables du joueur, mais piochés dans l'ensemble des contributions de tous les joueurs.

4 La suite

Le premier prototype est en attente de validation par le biais d'une évaluation qualitative des données récoltées lors d'une expérience de laboratoire. Celle-ci sera suivie d'une série de *playtests* destinés à perfectionner les éléments de la mécanique de jeu, ainsi qu'à mesurer l'intérêt du public cible. De nouvelles fonctionnalités sont déjà prévues pour les prototypes suivants, et les *playtests* serviront entre autres à recentrer le développement.

Le code source du jeu sera mis à disposition gratuitement sous licence libre dès la fin du test bêta public. Les ressources lexicales acquises par le biais du jeu seront de même publiées gratuitement sous licence ouverte.

Conversation libre Dans le prototype, le système sous-jacent ne tient pas encore compte des contributions du joueur, mais une version améliorée est actuellement en cours de développement.

Défis entre robots Les joueurs seront invités à participer à de petits défis opposant leurs robots au travers d'une série de problèmes à résoudre. Dans un premier temps, ces défis seront simples, et relativement proches des demandes d'A-LEX. Par la suite, des problèmes plus complexes seront considérés, en lien direct avec des problématiques de recherche.

Il est à noter que les joueurs sont des spectateurs passifs de ces défis, qui sont donc avant tout conçus pour renforcer l'immersion ludique, ainsi que servir de boucle de rétroaction principale pour les joueurs attirés par les jeux compétitifs. Ainsi, la seule possibilité pour un joueur compétitif d'avoir le « meilleur » robot est de passer du temps à l'aider à améliorer son vocabulaire.

Partage des connaissances Une variante des défis entre robots permettra à deux joueurs de mettre en commun leurs ressources pour résoudre une série de problèmes. La coopération est ici de mise et renforce l'immersion ludique pour les joueurs qui apprécient les jeux pour leurs aspects sociaux.

Visualisation du savoir Le joueur pourra inspecter le « cerveau » d'A-LEX à tout moment, et naviguer dans son réseau lexical pour étudier les effets de chaque interaction. Cette fonctionnalité tend plutôt à un usage éducatif du jeu, mais contribue aussi à renforcer l'immersion ludique.

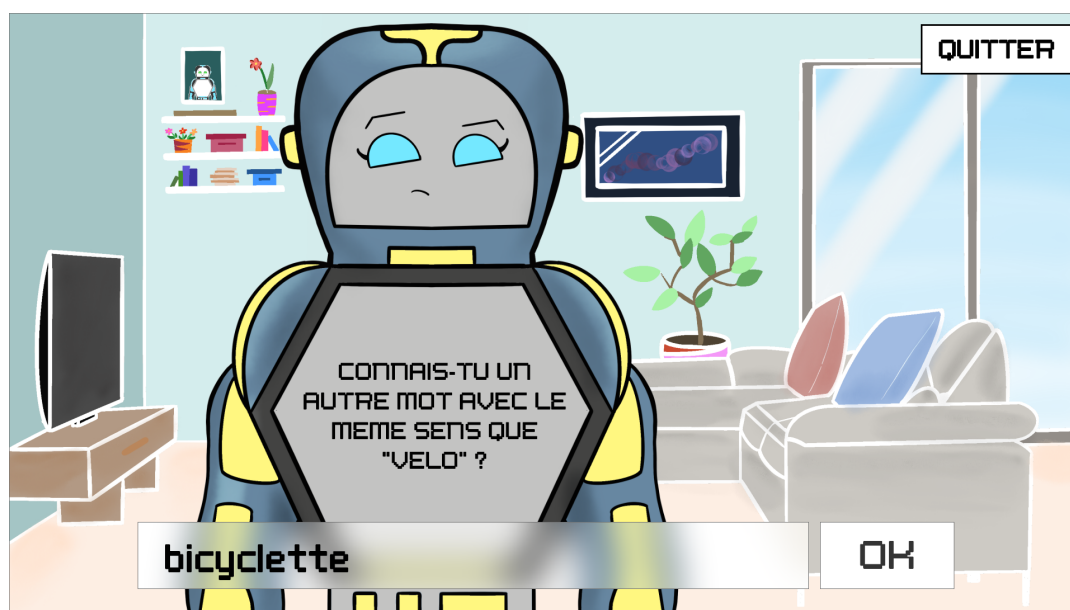


FIGURE 1 – Capture d'écran de la toute première version du prototype

Références

- ABT C. C. (1987). *Serious Games*. University Press of America. Original work published 1970.
- CSIKSZENTMIHALYI M. (2008). *Flow : The Psychology of Optimal Experience*. HarperCollins. Original work published 1990.
- LAFOURCADE M. & JOUBERT A. (2013). Bénéfices et limites de l'acquisition lexicale dans l'expérience JeuxDeMots. In N. GALA & M. ZOCK, Éd., *Ressources lexicales. Contenu, construction, utilisation, évaluation*, volume 30 de *Linguisticae Investigationes Supplementa*, p. 187–216. John Benjamins Publishing Company. DOI : [10.1075/lis.30](https://doi.org/10.1075/lis.30).
- MCGONIGAL J. (2011). We don't need no stinkin' badges : How to re-invent reality without gamification. In *GDC 2011 Serious Games Summit*, San Francisco, California, USA. <http://www.gdcvault.com/play/1014576/We-Don-t-Need-No>.
- SALEN K. & ZIMMERMAN E. (2003). *Rules of Play : Game Design Fundamentals*. Cambridge, Massachusetts : MIT Press.
- SCHRIER K. (2018). Designing learning with citizen science and games. *The Emerging Learning Design Journal*, 4(1). Source : <https://digitalcommons.montclair.edu/eldj/vol4/iss1/3>.
- SQUIRE K. (2008). Open-ended video games : A model for developing learning for the interactive age. In K. SALEN, Éd., *The Ecology of Games : Connecting Youth, Games, and Learning*, The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning, p. 167–198. Cambridge, MA : The MIT Press. DOI : [10.1162/dmal.9780262693646.167](https://doi.org/10.1162/dmal.9780262693646.167).
- TUITE K. (2014). GWAPs : Games with a problem. In *Proceedings of the 9th International Conference on the Foundations of Digital Games*, Ft. Lauderdale, FL, USA : Society for the Advancement of the Science of Digital Games. Source : http://www.fdg2014.org/papers/fdg2014_wip_22.pdf.
- VON AHN L. & DABBISH L. (2008). Designing games with a purpose. *Communications of the ACM*, 51(8), 58–67. Source : <http://doi.acm.org/10.1145/1378704.1378719>, DOI : [10.1145/1378704.1378719](https://doi.org/10.1145/1378704.1378719).

SkiF Maker : un processus de classement du meilleur talent par rapport à une offre de mission

Jihen Karoui¹ Kamilia Hafid¹ Stéphane Pezeril¹

(1) AUSY R&D, 6 rue Troyon, 92310 Sèvres Cedex, France

jkaroui@ausy.fr, kamilia.hafid@ausy.fr, spezeril@ausy.fr

RÉSUMÉ

SkiF Maker présente une solution innovatrice pour toute entreprise de services numériques (ESN) visant à améliorer son processus de recrutement de talents, la gestion des compétences de ses ressources humaines et la satisfaction des besoins de ses clients. Cet outil offre aux ESN un gain considérable en terme de temps de travail et leur évite un travail manuel souvent laborieux et non plaisant.

ABSTRACT

SkiF Maker : a process for ranking the best talent in relation to an assignment offer.

SkiF Maker presents an innovative solution for any digital services company aiming to improve its talent recruitment process, the management of its human resources skills and the satisfaction of the customer's requests. This tool provides digital service companies a considerable gain in terms of working time and avoids manual labor that is often laborious and unpleasant.

MOTS-CLÉS : Fouille de texte, annotation sémantique, apprentissage non supervisé, entité nommée.

KEYWORDS: Text mining, semantic annotation, unsupervised learning, named entity.

1 Contexte

Dans le métier du conseil en Hautes Technologies, le Curriculum-Vitae (CV) devient un outil du quotidien parce qu'il sert à mettre en avant les compétences d'un-e consultant-e. Le CV est un document commun, généralisé et incontournable dans le cadre d'un exposé de parcours professionnel mais il est loin d'avoir un contenu standardisé. Il est alors primordial d'adopter un format commun à tou-te-s, appelé "Dossier de Compétences (DC)". Le problème majeur est de trouver et d'ordonner les DC pour associer à l'offre de mission celui qui est le plus approprié. Appréhender un CV est donc une tâche complexe puisqu'il s'agit tant d'une analyse de structure de l'information sur une page que d'une langue.

L'évolution du marché du travail a prouvé que les méthodes classiques de recrutement ainsi que l'attribution des missions aux consultants sont devenues moins performant. Ceci est justifié par le développement considérable du recrutement en ligne ou l'e-recrutement ces dernières années (Faliagka *et al.*, 2014). Cette expansion a conduit à une croissance continue du nombre de portails de l'emploi et d'agences de recrutement sur Internet (Chen *et al.*, 2015), (Schmitt, 2016). Cela a également entraîné une augmentation constante du nombre de demandeurs d'emploi à la recherche de nouvelles opportunités de carrière (Hauff & Gousios, 2015). En conséquence, les portails d'emploi

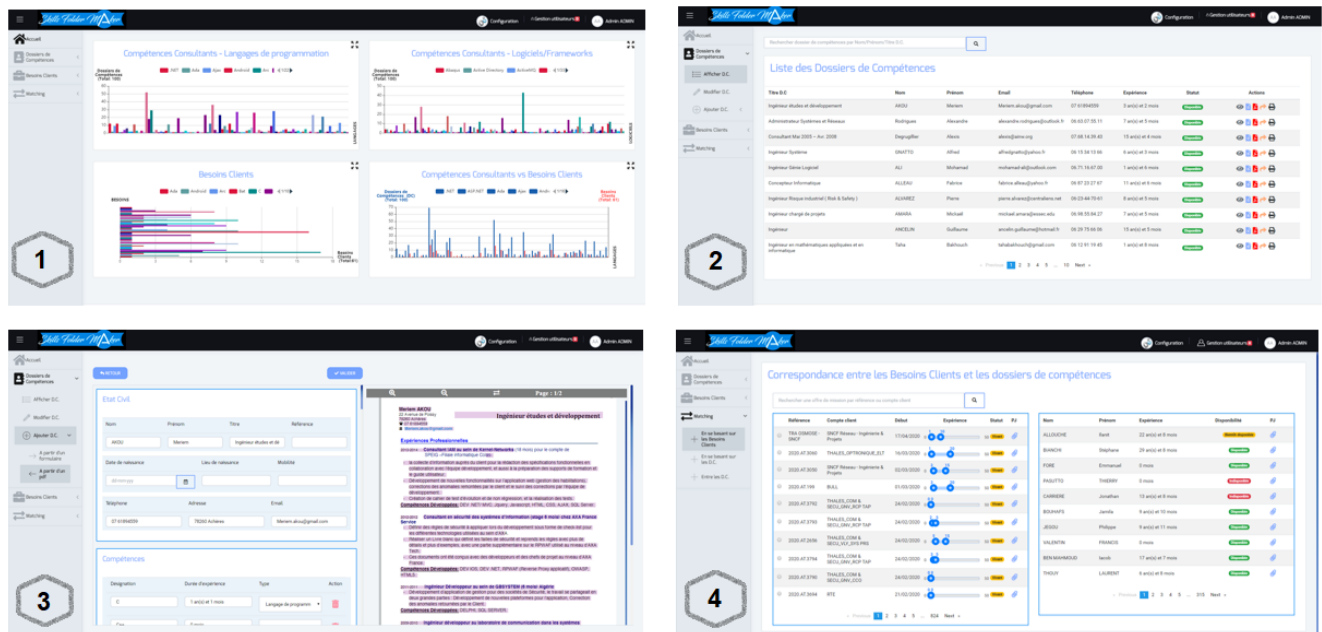


FIGURE 1 – Quelques interfaces graphiques de SkiF Maker

en ligne commencent à recevoir des milliers de CV (de styles et de formats divers) de demandeurs d'emploi possédant différents domaines d'expertise et spécialisés dans différents domaines (Kmail et al., 2015). L'analyse du contenu des CV ainsi que l'extraction des informations utiles pour la correspondance automatique entre les CV et les besoins clients n'a pas été résolu dans les différentes solutions présentes sur le marché aujourd'hui. Devant cette problématique, le développement d'une solution devient crucial et représente un challenge pour le TAL.

2 Le projet SkiF Maker

SkiF Maker est une solution innovatrice pour la gestion des compétences des salariés et des nouvelles recrues vs. les besoins clients. C'est un outil permettant aux conseillers en Hautes Technologies (voir figure 1) :

1. avoir une vision sur le marché à travers des dashboards qui présentent des statistiques sur les compétences des salariés et les compétences demandées par les clients (voir image N°1 de la figure 1).
2. la visualisation des données de l'entreprise relatives aux consultants (voir image N°2 de la figure 1).
3. l'extraction d'information et la génération automatique des DC à partir des CV des candidats (voir image N°3 de la figure 1).
4. la correspondance automatique entre les DC et les fiches de missions (voir image N°4 de la figure 1).

Dans le cadre de ce projet, nous avons dû faire face à plusieurs problématiques. La première problématique consiste à l'identification des différentes sections d'un CV ainsi que la **reconnaissance des entités nommés (EN)** présentant l'*état civil* (nom, prénom), les *compétences* (langages de programmation, logiciels, framework), les *formations* (intitulé des diplômes, écoles, universités), etc.

Cela nécessite l'exploitation des vocabulaires de domaine d'étude que nous avons construits. Pour la construction de ces vocabulaires, nous avons exploité une ressource existante à savoir **wiki-français**. Vu l'absence d'une ressource complète de vocabulaire spécifique au domaine de l'ingénierie informatique, nous avons opté pour un enrichissement manuel du vocabulaire issue de *wiki-français* dans un premier temps ensuite nous avons automatisé la veille technologique en interrogeant *Google* via son API afin de construire un vocabulaire spécifique au domaine de l'ingénierie informatique.

L'automatisation de la veille technologique s'effectue selon l'algorithme suivant. Considérons *WordsVocab* l'ensemble de mots excluant les mots vides qui sont présents dans la section compétences d'un CV. Soit *LP* (langage de programmation), *F* (framework) et *L* (logiciel) les catégories recherchées pour le classement du vocabulaire. L'algorithme proposé est défini comme suit :

Étapes de l'algorithme

1. Segmenter *WordsVocab* en un ensemble de tokens *T*.
 2. Générer trois requêtes pour chaque token :
 - (a) $Q1 = T \cup LP$
 - (b) $Q2 = T \cup F$
 - (c) $Q3 = T \cup L$
 soumettre à Google ces trois requêtes (une après l'autre) qui renverra pour chaque requête 20 résultats au maximum formés d'un titre et d'un extrait (snippet).
 3. A partir des résultats renvoyés par Google, ne garder que les deux premiers. Ensuite, pour chaque résultat, si les mots-clés de la requête Google ont été trouvés dans le titre ou dans l'extrait, alors *T* appartient à la catégorie trouvée (*LP* ou *F* ou *L*).
 4. Vérifier la présence de *T* dans le vocabulaire, si *T* n'existe pas alors il sera ajouter.
-

Cet algorithme nous a permis de construire un vocabulaire formé de 1200 termes. Pour évaluer l'exactitude de la catégorisation fourni par la veille technologique automatique, trois annotateurs ont également été invités à étiqueter l'ensemble des termes de *WordsVocab*. Les annotateurs ont confirmé l'exactitude des résultats fourni par l'algorithme.

Après avoir générer les DC et afin de résoudre la deuxième problématique, nous avons exploité un processus d'attribution de scores pour les compétences de chaque salarié/candidat. Ce processus consiste à calculer des scores pour l'ensemble des compétences présentes dans un DC en prenant en considération plusieurs critères. Parmi ces critères, nous citons le nombre d'année d'expérience global, le nombre d'année d'expérience pour chaque technos cité dans le DC, les compétences fonctionnelles effectués par le salarié durant ses expériences professionnelles précédentes (conception, développement, rédaction de la documentation, etc.). Cela a nécessité **l'analyse et l'annotation sémantique** du descriptif des expériences précédentes. Ces différents scores ont été exploités dans la phase d'**apprentissage non supervisé** et plus précisément le **Clustering**¹ afin que le système arrive à identifier et trier les DC qui correspondent à un besoin client donné et de retourner aux décideurs des ESN (manager, directeur, etc.) les meilleurs profils présents dans l'entreprise et les meilleurs profils à recruter.

1. Pour l'apprentissage, nous avons testé *K-means* avec deux mesures de similarité à savoir : la *distance Euclidienne* et la *distance de Manhattan*. Les premières expériences ont prouvé que la *distance de Manhattan* permet d'obtenir les meilleurs résultats.

Références

- CHEN J., NIU Z. & FU H. (2015). A novel knowledge extraction framework for resumes based on text classifier. In *International Conference on Web-Age Information Management*, p. 540–543 : Springer.
- FALIAGKA E., ILIADIS L., KARYDIS I., RIGOU M., SIOUTAS S., TSAKALIDIS A. & TZIMAS G. (2014). On-line consistent ranking on e-recruitment : seeking the truth behind a well-formed cv. *Artificial Intelligence Review*, **42**(3), 515–528.
- HAUFF C. & GOUSIOS G. (2015). Matching github developer profiles to job advertisements. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, p. 362–366 : IEEE Press.
- KMAIL A. B., MAREE M., BELKHATIR M. & ALHASHMI S. M. (2015). An automatic online recruitment system based on exploiting multiple semantic resources and concept-relatedness measures. In *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, p. 620–627 : IEEE.
- SCHMITT, T. P. C. M. S. (2016). Matching jobs and resumes : a deep collaborative filtering task. In *Proceedings of the 2nd Global Conference on Artificial Intelligence*, p. 1–14.

Iagotchi : vers un agent conversationnel artistique

Fréjus A. A. Laleye¹ Gaël de Chalendar¹ Léopold Frey² Rocio Berenguer²

(1) CEA, LIST, Laboratoire d'Analyse Sémantique Texte et Image, Gif-sur-Yvette, F-91191

(2) Compagnie Pulso, Marseille, 13001

frejus.laleye@cea.fr, gael.de-chalendar@cea.fr, leopold.frey@lilo.org,
creacionesinpulso@gmail.com

RÉSUMÉ

Cet article décrit Iagotchi, un personnage virtuel philosophique et artistique qui apprend et développe des connaissances à partir de ses interactions avec l'humain. Iagotchi se présente à la fois comme un apprenant et un expert avec comme objectifs principaux (1) d'accompagner l'homme dans ses questionnements, (2) de lui fournir des réponses pertinentes sur la base de ses requêtes et (3) de générer des textes poétiques cohérents. Dans ce travail, nous décrivons l'architecture du système de Iagotchi et les composants clés tels que le moteur de conversation, le gestionnaire de sujets et le générateur de poésies.

ABSTRACT

Iagotchi : Towards an Artistic Conversational Agent

This paper describes Iagotchi, a virtual philosophical and artistic character who learns and develops knowledge from his interactions. Iagotchi is designed as both a learner and an expert with the main objectives (1) to address human questions, (2) to provide with relevant responses on the basis of human requests and (3) to generate consistent poetic texts. In this work, we describe the Iagotchi system architecture and the key components such as conversation engine, topic manager and poetry generator.

MOTS-CLÉS : Iagotchi, Agents conversationnels, Chatbots, Art et Science.

KEYWORDS: Iagotchi, Conversationnal Agents, Chatbots, Art and Science.

1 Introduction

Dans ce travail, comme dans les récents travaux (Io & Lee, 2017; Gao *et al.*, 2019; Hussain *et al.*, 2019), nous avons combiné des méthodes traditionnelles du TALN et les approches neuronales pour concevoir un agent conversationnel qui lie l'art et la science. Iagotchi est un agent dialoguant *intelligent* capable de conduire une conversation de bout en bout avec l'humain. Il est conçu pour apprendre la pensée humaine et partager avec l'humain des intérêts intellectuels et des questions philosophiques telles que l'intelligence, le savoir, la vie, la pensée, l'amour etc. Il peut décider de répondre à une question sur la base de ses connaissances actuelles dans la mémoire de travail ou fournir une réponse significative à partir de sa mémoire sémantique. Une approche similaire a été utilisée dans (Bird *et al.*, 2019) pour détecter des messages offensants et s'est avérée efficace pour surmonter les problèmes rencontrés par des chatbots à domaine ouvert. Iagotchi a aussi la capacité de produire des poésies à la demande de son interlocuteur grâce à son moteur de génération de textes poétiques. En plus des questions philosophiques et de la poésie, il intègre des connaissances dans deux autres thématiques qui constituent les catégories principales de ses conversations : la *rencontre*

et le *G5 interespèces*. La catégorie *rencontre* regroupe l’ensemble des questions d’une conversation naturelle au cours de laquelle Iagotchi peut apprendre de son interlocuteur. Dans la catégorie *G5 interespèces*, Iagotchi se présente comme un expert des questions liées au G5¹.

2 Description de l’approche

Le fonctionnement de Iagotchi est fondé sur une approche modulaire qui tient compte du type de question et de réponse de l’humain. Il est aussi fondé sur une stratégie de gestion de dialogues mixte dans le sens où l’humain et Iagotchi peuvent tous deux conduire la conversation. Son architecture globale est présentée à la figure 1. Elle se compose de quatre couches :

- une interface audio ;
- un analyseur linguistique ;
- un moteur de conversation ;
- un ensemble de sources de données ;

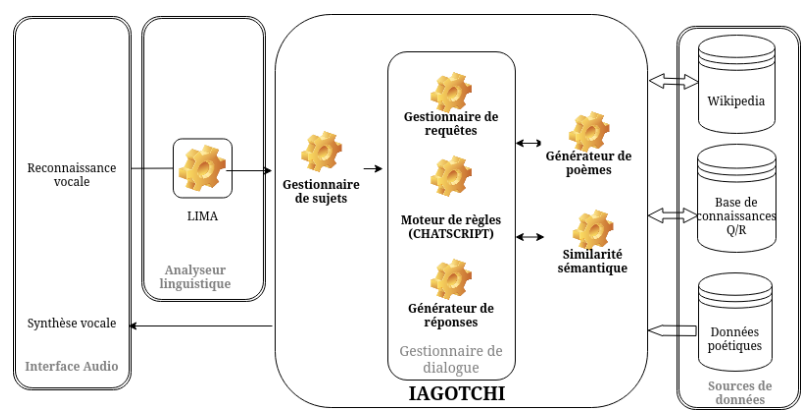


FIGURE 1 – Architecture du système

Algorithme 1 : Fonctionnement du gestionnaire de requêtes.

```

1  Considérant une question Q et une action A
2  suivant A faire
3      cas où A correspond à la recherche du sens de Q faire
4          |   RechercherSurWikipedia(Q) ;
5      fin
6      cas où A correspond à générer une poésie suite à Q faire
7          |   GénérerUnePoésie() ;
8      fin
9      autres cas faire
10         |   RetrouverQuestionSimilaire(Q) ;
11     fin
12 fin

```

2.1 Le gestionnaire de dialogues

Le gestionnaire de dialogues constitue l’élément central du moteur de conversations. Il est composé d’un ensemble de règles de dialogues écrites manuellement et de concepts définis à partir de l’en-

1. Le G5 est un forum politique informel de discussion entre les représentants des cinq règnes vivants sur terre : végétal, minéral, animal, humain et machine. Il a pour objectif d’identifier les mesures à prendre sur les grandes questions touchant à la mondialisation, d’impulser et de coordonner les décisions politiques correspondantes et de faciliter leur mise en oeuvre en liaison avec les organisations inter-espèces compétentes.

semble des données questions/réponses de base. Le fonctionnement du gestionnaire de requêtes et le processus d'interprétation d'une question sont respectivement décrits par les algorithmes 1 et 2.

Algorithme 2 : Fonctionnement du gestionnaire de dialogues de Iagotchi.

```

1  Considérant une question  $Q$  et un sujet  $S$ 
2  Substitution des mots par les concepts
3  Tri des règles selon la structure et les priorités
4  Chargement en mémoire des règles de  $S$ 
5  pour chaque règle de l'ensemble des règles en mémoire faire
6      si la règle correspond à  $Q$  alors
7          GénérerRéponse(action) ;
8              • renvoyer une réponse ;
9              • envoyer une requête vers Wikipedia ;
10             • générer un poème ;
11      sinon
12          RetrouverQuestionSimilaire( $Q$ ) ;
13              • envoyer  $Q$  au module de similarité ;
14      fin
15 fin
16 si une réponse est générée alors
17     Associer la réponse à  $Q$  en sauvegardant dans la base de connaissances Q/R
18 sinon
19     Sauvegarder  $Q$  dans la base de connaissances Q/R
20 fin

```

2.2 Le module de similarité sémantique

Ce module permet à Iagotchi, quand il n'a pas compris une question, de retrouver dans ses connaissances, celle qui est sémantiquement proche de la question de l'utilisateur. Il exploite les propriétés sémantiques capturées par les représentations distribuées de mots dans une phrase pour calculer une distance sémantique entre deux questions. Il intègre son propre générateur de réponse qui retourne la réponse R_c associée à la connaissance Q_c dont la distance par rapport à la question Q_u de l'utilisateur est proche de zéro. Nous avons utilisé la distance sémantique décrite dans (Laleye *et al.*, 2020). L'exemple (d) du dialogue dans le tableau 1 illustre l'utilisation du module de similarité pour répondre à une question incomprise par Iagotchi à l'aide de ses règles.

```

Humain : bonjour
Iagotchi: wouaoh, je t'attendais. bonjour comment tu t'appelles
Humain : je m'appelle Léo et toi comment ça va <-- (a)
Iagotchi: je vais bien. ok. comment t'expliques-tu les coïncidences?
Humain : je ne sais pas dit moi
Iagotchi: Pourquoi veux-tu parler de ça. Excuse moi je déconne un peu,
          je teste ta capacité à accepter l'incohérence, ça va?
Humain : c'est quoi le G5
Iagotchi: Le G5 est un forum politique informel de discussion entre les
          représentants des cinq règnes vivants sur terre. Organisé par <-- (b)
          le IOFLE.
Humain : c'est quoi l'intelligence
Iagotchi: L'art de la connerie avec science. <-- (c)
Humain : qu'est-ce qui se passe quand la connerie rencontre la science
Iagotchi: Pourquoi veux-tu parler de ça. Qu'est-ce qu'une rencontre ? <-- (d)
Humain : on se rencontre
Iagotchi: Ceci est une rencontre ?
Humain : je ne sais pas
Iagotchi: comment tu sais que tu sais quelque chose
Humain : au revoir
Iagotchi: ah tu pars déjà? bon merci quand même.

```

Étiquette	Thème	Action
(a)	Rencontre & G5	extraction de l'identité de l'humain & génération de réponse
(b)	G5	identification de règle & génération de réponse
(c)	Rencontre	identification de règle & génération de réponse
(d)	Rencontre	recherche de question similaire & génération de réponse

TABLE 1 — Un exemple de dialogue entre Iagotchi et un humain avec identification des types de processus mis en oeuvre

3 Mise en oeuvre à Expérimenta

Lors du salon Expérimenta à Grenoble, Iagotchi a été mis à disposition du public (Figure 2). Le



FIGURE 2 – Iagotchi, incarné par son cœur battant, discutant avec un visiteur

public d'Expérimenta est très varié : des scolaires, des chercheurs en sciences fondamentales et même en TAL, des artistes de disciplines diverses, du "grand public" et même du personnel politique... Iagotchi a globalement été très apprécié. Même si quelques personnes ressentaient une forte gêne à interagir avec une machine, la plupart éprouvaient une grande curiosité. Malgré les incohérences manifestes de quelques échanges, des personnes restaient discuter avec elle (ou lui) pendant plusieurs minutes, jusqu'à atteindre la durée maximale préétablie. Certain.e.s y apportaient manifestement toute leur volonté de dialogue. Citons quelques réactions à chaud d'adolescents : "C'est trop bien !", "Il comprend bien, c'est fou !".

4 Conclusion

Dans ce travail, nous avons présenté Iagotchi, un agent conversationnel conçu pour expérimenter un modèle relationnel gagnant-gagnant pour l'humain et la machine basé sur l'échange intellectuel et le questionnement philosophique. Iagotchi a été exposé pour des interactions avec le public afin de tester sa capacité intellectuelle, relationnelle et communicationnelle.

Références

- BIRD J. J., EKÁRT A. & FARIA D. R. (2019). Learning from interaction : An intelligent networked-based human-bot and bot-bot chatbot system. In *Advances in Computational Intelligence Systems*, p. 179–190, Cham : Springer International Publishing. DOI : https://doi.org/10.1007/978-3-319-97982-3_15.
- GAO J., GALLEY M. & LI L. (2019). Neural approaches to conversational ai. *Foundations and Trends® in Information Retrieval*, **13**(2-3), 127–298. DOI : [10.1561/15000000074](https://doi.org/10.1561/15000000074).
- HUSSAIN S., SIANAKI O. & ABABNEH N. (2019). A Survey on Conversational Agents/Chatbots Classification and Design Techniques, In *Primate Life Histories, Sex Roles, and Adaptability*, p. 946–956. DOI : [10.1007/978-3-030-15035-8_93](https://doi.org/10.1007/978-3-030-15035-8_93).
- IO H. N. & LEE C. B. (2017). Chatbots and conversational agents : A bibliometric analysis. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, p. 215–219. DOI : [10.1109/IEEM.2017.8289883](https://doi.org/10.1109/IEEM.2017.8289883).
- LALEYE F. A. A., BLANIÉ A., BROUQUET A., BENHAMOU D. & DE CHALENDAR G. (2020). Semantic similarity to improve question understanding in a virtual patient. In *TO APPEAR in Proceedings of the 35th ACM/SIGAPP Symposium On Applied Computing*.

Bien choisir son outil d'extraction de contenu à partir du Web

Gaël Lejeune¹ Adrien Barbaresi²

(1) Sorbonne Université, 1 rue Victor Cousin, 75005 Paris, France

(2) Académie des Sciences de Berlin-Brandenburg, Jägerstraße 22-23, 10117 Berlin, Allemagne

RÉSUMÉ

Nous proposons une démonstration sur l'extraction de contenu textuel dans des pages web ainsi que son évaluation. Nous nous concentrons sur les pages web contenant du texte (articles de presse, magazines en ligne et blogs) et montrons que les textes peuvent varier grandement selon différentes dimensions : diachronique, géographique et typologique. Dès lors, les outils et mesures d'évaluation correspondantes sont sujettes à caution : les indicateurs communément utilisés et censés présider au choix de l'outil approprié par les utilisateurs finaux sont à la fois imprécis et difficiles à interpréter.

ABSTRACT

Choosing the appropriate tool for Web Content Extraction

This demonstration focuses on the use and evaluation of Web Content Extraction tools, with a focus on web pages containing text (news articles, magazines and blogs). We show that the texts may differ with respect to diachronic, geographic and typological factors, so that state-of-the-art tools and measures are altogether imprecise and difficult to interpret.

MOTS-CLÉS : Construction de Corpus, Extraction de Contenu, Nettoyage de Pages Web, Cleaneval.

KEYWORDS: Web corpus construction, Web Content Extraction, Boilerplate removal, Cleaneval.

La construction de corpus à partir du Web comprend des opérations allant de la découverte et du téléchargement des sources jusqu'à l'annotation et l'indexation dans des bases de données (Kilgariff, 2007). Bien que cette construction soit devenue un élément commun des chaînes de traitement de TAL, les détails techniques concernant la mise en œuvre du nettoyage et de la préparation des pages web sont souvent omis. En définitive, chercheurs et utilisateurs s'en trouvent réduits à effectuer des évaluations a posteriori de la qualité et de l'adéquation des données (Baroni *et al.*, 2009). Le processus d'extraction de contenu peut être résumé de la façon suivante : étant donné le code source d'une page web, il s'agit d'extraire le contenu textuel utile et d'identifier les méta-données. Concrètement, cette tâche consiste notamment à écarter ce qui est de l'ordre du squelette de page web et implique une conversion du format HTML à un format texte ou XML. La disponibilité croissante d'outils génériques, par opposition à des approches ad hoc centrées sur l'application de règles d'extraction spécifiques à un site web particulier, a fait progressivement glisser cette tâche au rang d'outil d'ingénierie, alors même qu'elle a un impact direct sur les résultats scientifiques. D'autres approches, exploitant par exemple CommonCrawl¹ s'appuient sur une externalisation de la phase de *crawling* et l'extraction de contenu qui en découle (Habernal *et al.*, 2016). Nous laisserons de côté ici la question du choix des sources proprement dites pour nous concentrer sur les résultats de l'extraction, qui sont en eux-mêmes la base de décisions quant à l'inclusion d'un document donné dans le corpus final (Schäfer *et al.*, 2013). Peut-on réellement laisser de côté la question de ce que l'on a intégré dans des corpus et de l'impact sur les modèles qui vont en être extraits ? La nécessité pour

1. <https://commoncrawl.org>

les approches de type apprentissage profond de disposer de grandes quantités de données a forcément conduit à une plus grande légèreté sur la qualité ou la représentativité des données alors même que le besoin de recourir à une analyse fine existe toujours (Geyken *et al.*, 2017).

Nous montrons dans cette démonstration l’impact de différentes méthodes et outils d’extraction de contenus à partir du Web. En raison des performances affichées par les outils, a priori satisfaisantes et en progrès constants, l’impact du choix de l’outil sur la qualité peut être mésestimé par les utilisateurs finaux. Le fait qu’il y ait autant d’outils disponibles est en réalité un indicateur de la disparité dans la qualité des résultats obtenus. En ce sens, les métriques d’évaluation « état de l’art » proposent des scores agrégés qui laissent volontiers de côté des dimensions cruciales au profit d’une approche très générique et anglo-centrée (Barbaresi & Lejeune, 2020) et masquent trois dimensions :

1. Linguistique : les résultats sont extrêmement variables selon les pays d’origine et les langues
2. Typologique, quant à la nature et la forme des sites Web : les manières de construire une page web étant très variées, aucune garantie n’existe sur la robustesse des outils sur ce point
3. Diachronique : le langage du web évolue, comme en témoignent les standards et recommandations, si bien que les outils d’hier ne sont pas toujours adaptés aux pages d’aujourd’hui (Weninger *et al.*, 2016)

La dimension de la langue a souvent été escamotée alors que les (rares) expériences sur le sujet montrent une variabilité importante des résultats (Lejeune & Zhu, 2018). La dimension typologique a un impact en termes d’évaluation des besoins de l’utilisateur final : cherche-t-on un outil « tout terrain », efficace sur des sources variées, ou un outil « hi-fi », fiable sur les sources les plus fréquemment rencontrées ? Enfin, peut-on se fier aux résultats d’un outil sur des données de l’année X pour prédire son efficacité sur des données de l’année X+1 ? Cette dimension ne nous semble pas avoir obtenu l’attention qu’elle mérite dans la littérature sur les corpus web. Les « bonnes » propriétés qui assuraient de bons résultats à un temps T sont-elles conservées au fil de l’adaptation d’un outil aux données nouvelles ? La rétro-comparabilité des données² est-elle assurée ? Nous présentons ici des outils d’extraction de contenu parmi les plus utilisés avec un éclairage sur leurs performances, avec pour référence l’environnement Python, très présent si ce n’est majoritaire dans le monde de la recherche³. Certains outils étant adaptés d’autres langages, notre comparatif permet d’offrir un large tour d’horizon. Dans le tableau 1 nous reprenons la catégorisation de (Barbaresi & Lejeune, 2020).

Cat.	Outil	Version	Adresse Github	Référence
I	HTML2TEXT	2020.1.16	Alir3z4/html2text/	
I	INSCRIPTIS	1.0	weblyzard/inscriptis	
II	NEWSPAPER3K	0.2.8	codelucas/newspaper	
II	NEWS-PLEASE	1.4.25	fhamborg/news-please	(Hamborg <i>et al.</i> , 2017)
II	READABILITY	0.7.1	buriy/python-readability	
III	BOILERPY3	1.0.2	jmriebold/BoilerPy3	(Kohlschütter <i>et al.</i> , 2010)
III	DRAGNET	2.0.4	dragnet-org/dragnet	(Peters & Lecocq, 2013)
III	GOOSE3	3.1.6	goose3/goose3	
III	JUSTEXT	2.2.0	miso-belica/jusText	(Pomikálek, 2011)
III	TRAFILATURA	0.4.1	adbar/trafilatura	(Barbaresi, 2019)

TABLE 1: Outils orientés rappel (I), orientés lisibilité(II), spécifiquement dédiés à la tâche (III)

2. Par analogie avec la rétro-compatibilité.

3. <https://spectrum.ieee.org/computing/software/the-top-programming-languages-2019>

	multi	el	en	pl	ru	zh	Tps (sec.)	Diff/réf
TRAFILATURA_FB	75,69	81,29	84,30	75,13	68,51	69,22	109,9	x5,6
READABILITY	74,62	84,81	85,24	76,08	71,79	55,2	56,8	x2,9
BOILERPY3_ART	72,73	63,06	82,24	80,41	63,02	74,91	39,8	x2,0
JUSTEXT	63,7	86,55	80,33	78,97	70,47	2,18	322,0	x16,3
JUSTEXT_LANGID	63,31	86,74	79,26	78,56	69,83	2,18	112,6	x5,7
DRAGNET	58,21	34,0	86,04	72,81	43,31	54,89	24,0	x1,2
NEWSPLEASE	48,83	47,93	86,71	78,05	27,26	4,22	3755,6	x190
INSCRIPTIS	40,10	48,48	43,37	40,1	29,6	38,94	19,7	x1
GOOSE	37,87	2,45	88,6	68,26	27,0	3,06	191,3	x9,7
NEWSPAPER	32,37	3,74	89,3	63,34	3,37	2,11	105,5	x5,5
HTML2TEXT	31,2	38,43	41,64	32,27	26,37	17,31	71,0	x3,6
JUSTEXT_EN	17,63	2,11	79,26	1,71	2,96	2,11	41,5	x2,1

TABLE 2: Résultats sur le corpus DANIEL, Macro-moyenne des F-mesures sur la classe langue, et f-mesure par langue (mesure CLEANVAL), les temps de calcul sont une moyenne sur 5 tests et sont exprimés en secondes, le ratio de vitesse est calculé par rapport à l’outil le plus rapide (INSCRIPTIS)

Les comparaisons détaillées de ces outils sont accessibles dans (Barbaresi & Lejeune, 2020) et sont reprises sur un dépôt GITHUB dédié⁴ qui doit permettre non seulement la reproductibilité de l’évaluation mais également son suivi, certains outils étant encore activement en développement. Nous reprenons dans le tableau 2 une partie des résultats afin de montrer à quel point les outils peuvent avoir des comportements très variables en termes d’efficacité générale, en termes d’efficacité par langue et enfin en termes de temps de traitement, ce qui n’est pas négligeable. L’évaluation présentée ici utilise la F-mesure obtenue avec les mesures de CLEANVAL (Baroni *et al.*, 2008), elle a été menée sur le corpus DANIEL (Lejeune *et al.*, 2012; Lejeune & Zhu, 2018) qui comporte un peu plus de 1600 documents en 5 langues avec leur version HTML et leur version nettoyée manuellement.

Nous pouvons voir que les différences de résultats entre les outils sont significatives et que la qualité des résultats obtenus sur l’anglais ne préjuge absolument pas de la consistance des résultats concernant d’autres langues. Pour illustrer ce point nous avons fait figurer dans le tableau le résultat du modèle JUSTEXT sur l’anglais. À la vue de ce comparatif, il ne semble pas exister de solution « tout terrain » même si l’on observe un certain parallélisme entre l’anglais et le polonais. En outre, la vitesse n’est pas une dimension à négliger puisque si l’on compare au système le plus rapide (INSCRIPTIS) on observe des changements d’ordre de grandeur crucial pour l’abord de données massives : avec GOOSE, le résultat est presque 10 fois plus lent, plus de 15 fois plus lent pour le modèle indépendant de la langue de JUSTEXT. Ce dernier cas est intéressant puisque ce modèle indépendant est légèrement plus efficace que le modèle `langid` mais au prix d’un temps de traitement 3 fois plus grand. Enfin, NEWSPLEASE est un cas particulier puisque c’est un outil qui fait bien d’autres choses que le nettoyage.

Bien choisir son outil d’extraction de contenu passe donc non seulement par le choix et la configuration des outils en fonction des sources, mais aussi par des comparatifs et des observations, même à échelle réduite, qui permettent d’exploiter les différences tant sur l’efficacité des extracteurs que sur les temps de traitement.

4. <https://github.com/rundimeco/waddle>

Références

- BARBARESI A. (2019). Generic Web Content Extraction with Open-Source Software. In *Proceedings of KONVENS 2019, Kaleidoscope Abstracts*, p. 267–268 : GSCL.
- BARBARESI A. & LEJEUNE G. (2020). Out-of-the-Box and Into the Ditch ? Multilingual Evaluation of Generic Text Extraction Tools. In *Proceedings of the 12th Web as Corpus workshop (WAC-XII)* : ELRA. à paraître.
- BARONI M., BERNARDINI S., FERRARESI A. & ZANCHETTA E. (2009). The WaCky Wide Web : a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, **43**(3), 209–226.
- BARONI M., CHANTREE F., KILGARRIFF A. & SHAROFF S. (2008). Cleaneval : a Competition for Cleaning Web Pages. In *Proceedings of LREC*, p. 638–643 : ELRA.
- GEYKEN A., BARBARESI A., DIDAKOWSKI J., JURISH B., WIEGAND F. & LEMNITZER L. (2017). Die Korpusplattform des "Digitalen Wörterbuchs der deutschen Sprache" (DWDS). *Zeitschrift für germanistische Linguistik*, **45**(2), 327–344.
- HABERNAL I., ZAYED O. & GUREVYCH I. (2016). C4Corpus : Multilingual Web-size corpus with free license. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 914–922.
- HAMBORG F., MEUSCHKE N., BREITINGER C. & GIPP B. (2017). news-please : A generic news crawler and extractor. In M. GAEDE, V. TRKULJA & V. PETRA, Éd., *Proceedings of the 15th International Symposium of Information Science*, p. 218–223.
- KILGARRIFF A. (2007). Googleology is bad science. *Computational Linguistics*, **33**(1), 147–151.
- KOHLSCHÜTTER C., FANKHAUSER P. & NEJDL W. (2010). Boilerplate detection using shallow text features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM '10*, p. 441–450.
- LEJEUNE G., BRIXTTEL R., DOUCET A. & LUCAS N. (2012). Daniel : Language independent character-based news surveillance. In *International Conference on NLP*, p. 64–75 : Springer.
- LEJEUNE G. & ZHU L. (2018). A New Proposal for Evaluating Web Page Cleaning Tools. *Computación y Sistemas*, **22**(4).
- PETERS M. E. & LECOCQ D. (2013). Content extraction using diverse feature sets. In *Proceedings of the 22nd International Conference on World Wide Web*, p. 89–90.
- POMIKÁLEK J. (2011). *Removing boilerplate and duplicate content from web corpora*. Thèse de doctorat, Masaryk University.
- SCHÄFER R., BARBARESI A. & BILDHAUER F. (2013). The Good, the Bad, and the Hazy : Design Decisions in Web Corpus Construction. In *Proceedings of the 8th Web as Corpus Workshop*, p. 7–15.
- WENINGER T., PALACIOS R., CRESCENZI V., GOTTRON T. & MERIALDO P. (2016). Web Content Extraction – a Meta-Analysis of its Past and Thoughts on its Future. *ACM SIGKDD Explorations Newsletter*, **17**(2), 17–23.

Conception d'un système de détection d'intention pour un moteur de recherche sur Internet

Estelle Maudet Christophe Servan

Qwant Research, 7 Rue spontini, 75116 Paris, France

intial.lastname@qwant.com

RÉSUMÉ

Dans les moteurs de recherche sur Internet, l'une des tâches les plus importantes vise à identifier l'intention de l'utilisateur. Cet article présente notre étude pour proposer un nouveau système de détection d'intention pour le moteur de recherche sur Internet Qwant. Des logs de clic au système de détection d'intention, l'ensemble du processus est expliqué, y compris les contraintes industrielles qui ont dû être prises en compte. Une analyse manuelle des données groupées a d'abord été appliquée sur les journaux afin de mieux comprendre les objectifs de l'utilisateur et de choisir les catégories d'intention pertinentes. Lorsque la recherche satisfait aux contraintes industrielles, il faut faire des choix architecturaux et faire des concessions. Cet article explique les contraintes et les résultats obtenus pour ce nouveau système en ligne.

ABSTRACT

Designing a User Intention Detection system for a Web Search Engine

In web search engines, one of the most important tasks aims to identify the user's intention. This paper presents our study to propose a new intention detection system for the Qwant web search engine. From the click logs to the detection server, the entire process is explained, including the industrial constraints that had to be taken into account. A manual analysis of clustered data was first applied on the logs to better understand the user's goals and choose relevant intent categories. When research meets industrial constraints, some architectural choices and concessions have to be made. This paper explains the constraints and the results obtained for this new online system.

MOTS-CLÉS : Détection d'intention, Classification, humain-dans-la-boucle, extraction d'information, Recherche industrielle.

KEYWORDS: Intention Detection, Classification Task, Human-in-the-loop Clustering, Information Retrieval, Industrial research.

1 Introduction

Une intention est un but derrière une action spécifique ou un ensemble d'actions. Lorsqu'un utilisateur fait une requête sur un moteur de recherche, il a généralement un but spécifique qui peut être identifié et classé. Dans ce contexte, la détection d'intention est une caractéristique clé d'un moteur de recherche sur Internet.

Au fur et à mesure que la recherche sur les objectifs de l'utilisateur progresse, il est nécessaire de caractériser plus précisément les requêtes afin de mieux répondre aux besoins de l'utilisateur (Baeza-Yates *et al.*, 2006). Au début, des systèmes fondés sur des règles ont été utilisés pour identifier des tentatives particulières. Des études plus récentes ont favorisé l'utilisation de solutions d'apprentissage

automatique en utilisant des logs de clics et des données Internet supplémentaires comme Wikipedia (Gabrilovich *et al.*, 2009; Ren *et al.*, 2014; Hashemi *et al.*, 2016). Depuis lors, la tâche de détection d'intention peut être considérée comme une tâche de classification. Plusieurs études ont été menées au cours des dernières années avec succès (Kim, 2014; Lai *et al.*, 2015; Zhang *et al.*, 2015; Conneau *et al.*, 2017). Toutefois, le principal inconvénient de ces approches est le non-respect des contraintes industrielles.

Ce travail a eu lieu dans un contexte industriel, le moteur de recherche Qwant. Elle implique des contraintes spécifiques en raison de la spécificité du moteur de recherche ainsi que des impératifs de production. Le premier défi industriel est le passage à l'échelle du système de détection. Comme les utilisateurs n'aiment pas attendre pour avoir leur réponse, la détection de l'intention doit avoir un faible impact sur la latence globale du moteurs de recherche. Le deuxième défi industriel était l'absence de métadonnées disponibles pour contextualiser la requête. En fait, Qwant est un moteur de recherche sur Internet axé sur la confidentialité, ce qui signifie qu'aucune information personnelle n'est stockée sur aucun serveur ni collectée (pas d'historique, pas d'adresse IP, pas de session). Seuls les logs de clics peuvent être utilisés dans le modèle d'intention. Ces logs se composent d'une liste de tuples avec une requête faite par un utilisateur anonyme et l'URL qui a été cliquée.

2 Collecte, regroupement et tri des données

L'analyse des regroupements (ou grappes) joue un rôle important dans le domaine de l'exploitation des données. Dans cette étude, l'algorithme des K-Moyennes a été utilisé sur les requêtes comme tâche préliminaire pour trouver de l'information sur les modèles. Nous avons obtenu des regroupements fins avec $K = 500$.

Alors que Wikipedia est écrit en langue naturelle avec peu d'erreurs d'orthographe, les requêtes des logs sont très bruitées et très courtes (la longueur moyenne d'une requête est d'environ 3,1 mots). Nous regroupons les requêtes à l'aide de représentation continues de phrases en considérant ces dernières comme un sac de mots (Joulin *et al.*, 2017). Nous avons appris des représentations continues de mots de 300 dimensions sur l'ensemble des données (le Wikipedia français et les logs de clics complets). Utilisant les résultats du regroupement automatique, chaque grappe a été traitée manuellement pour identifier les intentions pertinentes et les regrouper en utilisant les sites Internet les plus pertinents comme *graine*. Nous avons traité de la même façon pour toutes les classes et obtenu une liste de *graines* intéressantes pour chaque intention. (p.ex. : *jeuxvideo.com*, *gamekult.com* et *steampowered.com* pour la classe « Jeux Vidéos »).

Enfin, 100 millions de requêtes ont été regroupées en seize catégories pour créer le modèle de classification d'intention. 2000 requêtes additionnelles ont été annotées manuellement pour l'évaluation.

3 Performances & conclusion

Le modèle obtenu a un score de 83,60 de précision en considérant les 16 classes du corpus de test. Afin de répondre aux contraintes industrielles, le modèle a été chargé à l'aide d'un serveur REST en C++ *Pistache*¹. The temps de réponse moyen est de 0,45 ms, ce qui représente environs 2 200 requêtes par seconde. Le code de l'API est disponible sur GitHub². Enfin, le déploiement a permis une amélioration mesurée du taux de clic sur le *Shopping* de plus de 35%.

1. *Pistache* (pistache.io)

2. <https://github.com/QwantResearch/text-classifier/>

Références

- BAEZA-YATES R., CALDERÓN-BENAVIDES L. & GONZÁLEZ-CARO C. (2006). The intention behind web queries. In *International Symposium on String Processing and Information Retrieval*, p. 98–109 : Springer.
- CONNEAU A., SCHWENK H., BARRAULT L. & LECUN Y. (2017). Very Deep Convolutional Networks for Text Classification. In *Proceedings of the 15th Conference of the {E}uropean Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, p. 1107–1116, Valencia, Spain : Association for Computational Linguistics.
- GABRILOVICH E., BRODER A., FONTOURA M., JOSHI A., JOSIFOVSKI V., RIEDEL L. & ZHANG T. (2009). Classifying search queries using the web as a source of knowledge. *ACM Transactions on the Web (TWEB)*, **3**, 5.
- HASHEMI H. B., ASIAEE A. & KRAFT R. (2016). Query intent detection using convolutional neural networks. In *International Conference on Web Search and Data Mining, Workshop on Query Understanding*.
- JOULIN A., GRAVE E., BOJANOWSKI P. & MIKOLOV T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, p. 427–431.
- KIM Y. (2014). Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing ({EMNLP})*, p. 1746–1751, Doha, Qatar : Association for Computational Linguistics. DOI : [10.3115/v1/D14-1181](https://doi.org/10.3115/v1/D14-1181).
- LAI S., XU L., LIU K. & ZHAO J. (2015). Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.
- REN X., WANG Y., YU X., YAN J., CHEN Z. & HAN J. (2014). Heterogeneous graph-based intent learning with queries, web pages and wikipedia concepts. In *Proceedings of the 7th ACM international conference on Web search and data mining*, p. 23–32 : ACM.
- ZHANG X., ZHAO J. & LECUN Y. (2015). Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, p. 649–657.

Projet AMIS : résumé et traduction automatique de vidéos

Mohamed Amine Menacer Dominique Fohr Denis Juvet Karima Abidi
David Langlois Kamel Smaïli

Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France
prenom.nom@loria.fr

RÉSUMÉ

La démonstration de résumé et de traduction automatique de vidéos résulte de nos travaux dans le projet AMIS. L'objectif du projet était d'aider un voyageur à comprendre les nouvelles dans un pays étranger. Pour cela, le projet propose de résumer et traduire automatiquement une vidéo en langue étrangère (ici, l'arabe). Un autre objectif du projet était aussi de comparer les opinions et sentiments exprimés dans plusieurs vidéos comparables. La démonstration porte sur l'aspect résumé, transcription et traduction. Les exemples montrés permettront de comprendre et mesurer qualitativement les résultats du projet.

ABSTRACT

AMIS project : automatic summarization and translation of videos

The demonstration of video summarization and machine translation is the result of our work in the AMIS project. The project aimed at helping a traveller to understand news in a foreign country. For that, the solution was to provide summaries and translations of videos given in a foreign language (here Arabic). Another goal of the project was to compare sentiments and opinions expressed in comparable videos. to compare in terms of sentiments and opinions comparable videos. During the demo, we will present several videos in Arabic, and their translation in English. The given examples will be representative of the results of the project.

MOTS-CLÉS : vidéo, résumé automatique, traduction automatique, arabe, anglais.

KEYWORDS: video, automatic summarization, machine translation, Arabic, English.

1 Introduction

Imaginons un voyageur dans une ville à l'étranger. Un événement local survient, qui fait la une des journaux télévisés. Le voyageur passe d'une chaîne à une autre, mais ne comprend que partiellement ce qu'il se passe. Pourtant, cela peut avoir un impact fort sur son séjour (fermeture d'aéroport, limitation de circulation). Le voyageur a donc besoin, depuis son hôtel, d'avoir un point rapide sur la situation dans sa propre langue.

Concrètement, répondre à ce besoin implique plusieurs défis scientifiques : extraire un résumé d'une vidéo en se basant sur l'information vidéo (résumé de vidéo) et sonore (résumé de texte), et traduire le flux sonore dans une autre langue (traduction parole-texte, ou même parole-parole). Un autre défi est la recherche du flux de traitement le plus performant, sachant que chaque étape entraîne des erreurs : comment tenir compte des critères de résumé vidéo (rendre compte des différents lieux filmés, et peut-être ne pas insister trop longtemps sur le présentateur du journal) et des critères de résumé textuel

(tenir compte des idées exprimées) et marier ces critères ? Doit-on transcrire et traduire le résumé, ou bien transcrire, traduire la totalité de la vidéo, puis la résumer, etc. ? Toute la combinatoire est possible.

Le projet AMIS¹ a cherché à répondre à ces questions. Au-delà, un objectif était aussi de permettre à l'utilisateur de comprendre les différents points de vue via la comparaison en termes de sentiments et d'opinions de plusieurs vidéos comparables ; mais la démonstration ne concernera que les aspects de traduction et résumé. Ce projet de type Chist-Era implique plusieurs partenaires internationaux experts en extraction d'information vidéo (AGH², Cracovie, Pologne), en résumé de texte (LIA³, Avignon, France), en transcription et traduction automatique de parole (Loria⁴, Nancy, France), en test utilisateur (Deusto⁵, Bilbao, Espagne). Nous décrivons ci-après les différentes configurations de résumé-vidéo-texte-traduction (Section 2) issues du projet, puis nous présentons la démonstration que nous montrerons pendant la conférence (Section 3).

2 Le processus de résumé/traduction

Le projet a abouti à quatre scénarios de fabrication d'un résumé vidéo dans une autre langue. Les scénarios reposent sur l'utilisation de la reconnaissance de la parole, de la traduction de texte, et sur diverses approches de résumé automatique : à partir des images, de l'audio ou du texte. Ces différents composants ont d'abord été évalués individuellement sur la base des métriques d'évaluation usuelles (Smaïli *et al.*, 2019). Puis les scénarios ont été comparés sur la base de retours utilisateur concernant la compréhensibilité du résumé, et son adéquation par rapport à la vidéo source.

Les quatre scénarios mis en œuvre et schématisés sur la figure 1 sont les suivants :

- Sc1 repose sur un résumé à partir des images vidéo. L'audio des segments sélectionnés est ensuite traité par la reconnaissance de la parole, et le résultat est traduit dans la langue cible.
- Sc2 repose sur un résumé à partir des informations audio. Les segments obtenus sont ensuite traités par la reconnaissance de la parole et traduits dans la langue cible.
- Sc3 repose sur la reconnaissance de parole de l'émission complète et sa traduction dans la langue cible. Le résumé est appliqué sur le texte traduit. Enfin les segments vidéo correspondant aux phrases ou portions de phrase du résumé sont extraits et concaténés pour fournir le résumé de l'émission.
- Sc4 est similaire à Sc3, mais le résumé est réalisé sur le résultat de la reconnaissance de parole, puis traduit dans la langue cible.

3 La démonstration

Au cours de la démonstration, nous montrerons des vidéos en langue arabe, et leur résumé en langue anglaise (voir Figure 2). Du fait des temps de calcul non temps réel, le processus de résumé/traduction

1. Access Multilingual Information opinionS, <http://deustotechlife.deusto.es/amis/>

2. Akademia Górniczo-Hutnicza, <http://deustotechlife.deusto.es/amis/partner/3>

3. Laboratoire d'Informatique d'Avignon, <http://deustotechlife.deusto.es/amis/partner/5>

4. Laboratoire Lorrain en Informatique et ses Applications, <http://deustotechlife.deusto.es/amis/partner/2>

5. Université de Deusto, <http://deustotechlife.deusto.es/amis/partner/4>

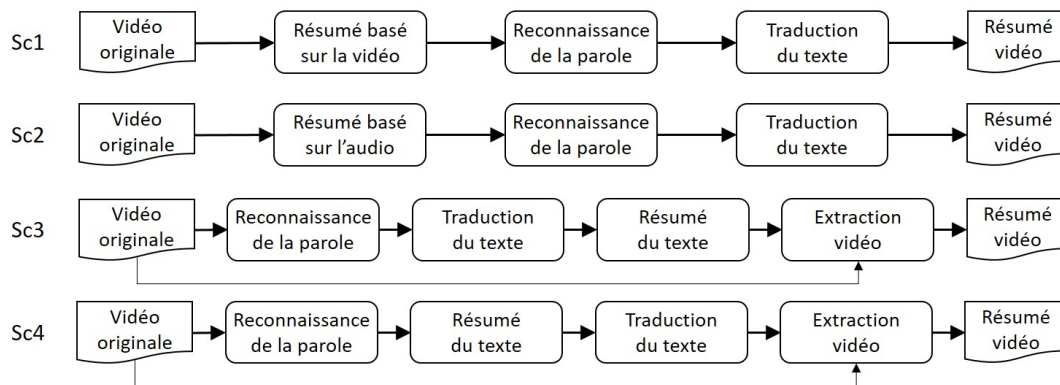


FIGURE 1 – Architectures pour la fabrication de résumés de vidéos dans une autre langue.

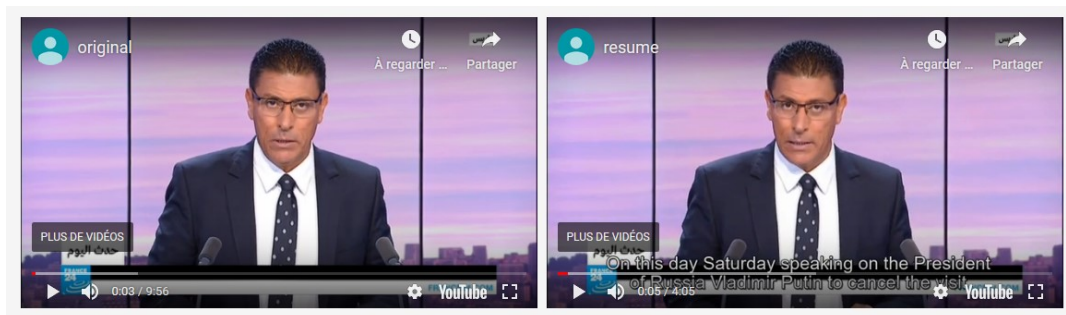


FIGURE 2 – Capture d’écran de la démonstration, à gauche la vidéo initiale en arabe (10 minutes), à droite la vidéo résumée (4 minutes) et les sous-titres en anglais.

ne se fera pas en direct. La traduction sera donnée sous forme de sous-titres. Le contenu arabophone des vidéos sera décrit. La démonstration sera l’occasion de relever les passages qui ont été bien sélectionnés et traduits, et ceux pour lesquels les résultats sont à améliorer. Les différentes vidéos montreront des résultats suite aux traitements des quatre scénarios. La démonstration sera aussi l’occasion de revenir sur les résultats publiés (Menacer *et al.*, 2019; Smaïli *et al.*, 2018), et de les commenter sur des exemples, et enfin de répondre à toute question sur les aspects scientifiques et techniques des méthodes utilisées pour la transcription et la traduction.

Remerciements

Nous remercions Chist-Era pour avoir financé ce travail via le projet AMIS (Access Multilingual Information opinionS). Cette démonstration a été organisée par le LORIA, mais se base sur les travaux de l’ensemble des partenaires AGH, DEUSTO, LIA, et LORIA.

Références

MENACER M. A., GONZÁLEZ-GALLARDO C. E., ABIDI K., FOHR D., JOUVET D., LANGLOIS D., MELLA O., SADAT F., TORRES-MORENO J. M. & SMAÏLI K. (2019). Extractive Text-Based Summarization of Arabic videos : Issues, Approaches and Evaluations. In *ICALP : International*

Conference on Arabic Language Processing, volume Communications in Computer and Information Science book series (CCIS, volume 1108), p. 65–78, Nancy, France : Springer.

SMAÏLI K., FOHR D., GONZÁLEZ-GALLARDO C., GREGA M., JANOWSKI L., JOUVET D., KOMOROWSKI A., KOZBIAL A., LANGLOIS D., LESZCZUK M., MELLA O., MENACER M. A., MENDEZ A., LINHARES PONTES E., SANJUAN E., SWIST D., TORRES-MORENO J.-M. & GARCIA-ZAPIRAIN B. (2018). A First Summarization System of a Video in a Target Language. In *MISSI 2018 - 11th edition of the International Conference on Multimedia and Network Information Systems*, p. 1–12, Wroclaw, Poland.

SMAÏLI K., FOHR D., GONZÁLEZ-GALLARDO C.-E., GREGA M. L., JANOWSKI L., JOUVET D., KOŻBIAL A., LANGLOIS D., LESZCZUK M., MELLA O., MENACER M.-A., MENDEZ A., PONTES E. L. L., SANJUAN E., TORRES-MORENO J.-M. & GARCIA-ZAPIRAIN B. (2019). Summarizing videos into a target language : Methodology, architectures and evaluation. *Journal of Intelligent and Fuzzy Systems*, **1**, 1–12.

La résolution d'anaphores au-delà de la frontière de la phrase

Luka Nerima, Eric Wehrli

Laboratoire d'analyse et de technologie du langage (LATL)

CUI - Université de Genève, 7 route de Drize, CH-1227 Carouge, Suisse

luka.nerima@unige.ch, eric.wehrli@unige.ch

RÉSUMÉ

Cette démonstration présente une extension de nos outils d'analyse syntaxique et d'étiquetage morphosyntaxique qui prend en compte la résolution d'anaphores pronominales non seulement à l'intérieur d'une phrase, mais également si l'antécédent se trouve dans la phrase précédente. Autant l'analyseur que l'étiqueteur effectuant une analyse syntaxique complète des phrases, ces outils affichent également les fonctions grammaticales des constituants (sujet, objet direct, etc.) et les arguments des verbes. Une version de cette démonstration est disponible sur le Web.

ABSTRACT

The Anaphora Resolution Beyond Sentence Boundary

This demo presents an extension of our syntactic analysis and morphosyntactic tagging tools that takes into account the resolution of pronominal anaphors not only within a single sentence, but also if the antecedent is in the previous sentence. A version of this demo is available on the Web.

MOTS-CLES : Résolution d'anaphores, analyseur syntaxique, étiquetage morphosyntaxique.

KEYWORDS : Anaphora resolution, natural language parsing, POS tagging.

1 Introduction

Malgré les avancées spectaculaires sur les différents fronts des applications de TAL ces 5 dernières années, la résolution d'anaphores (RA) n'a pas progressé de manière significative. Les exemples de tâches où cette résolution joue un rôle important sont pourtant nombreux : traduction automatique, résumés automatiques, reconnaissance d'entités nommées, extraction et identification d'expressions polylexicales, etc. (voir p. ex. Mitkov, 2002, Sukthankar 2018). L'étiqueteur morphosyntaxique que nous montrons ici prend en charge une résolution d'anaphore particulière, celle des antécédents des pronoms personnels de la troisième personne. À noter aussi que, dans leur version actuelle, nos outils d'analyse ne traitent pas les cataphores.

2 La méthode

Dans le cadre de la grammaire générative, la théorie du liage de Chomsky (1981) et les implémentations informatiques de Hobbs (1978) et de Lappin et Leass (1994) sont sans doute les contributions les plus significatives à la RA. La théorie de la liaison de Chomsky n'est pas destinée à

être une méthode de RA en soi, mais un ensemble de contraintes sur la distribution des pronoms réfléchis et/ou réciproques (appelés anaphores) et des pronoms référentiels (appelés pronoms) qui sont extrêmement utiles pour filtrer l'ensemble des antécédents potentiels.

En bref, la théorie du liage stipule (i) que les pronoms réfléchis et/ou réciproques doivent être liés dans leur catégorie minimale de gouvernance, et (ii) que les pronoms référentiels ne peuvent pas être liés dans leur catégorie minimale de gouvernance, où "lié" signifie gouverné et co-indexé. En résumé, nous pouvons définir la catégorie minimale de gouvernance d'un élément comme le constituant minimal qui l'inclut. Il découle des principes du liage (i) et (ii) que les anaphores et les pronoms (pour reprendre la terminologie de Chomsky) ont une distribution complémentaire.

3 La démonstration

Dans les grandes lignes, la mise en œuvre de la RA suit l'algorithme de Lappin et Leass (1994), adapté aux spécificités des représentations grammaticales de notre analyseur. Appliquée à l'exemple de la Figure 1, la première tâche vérifie que le pronom *il* (à la ligne 3 de la deuxième phrase) n'est pas impersonnel (comme p. ex. dans *il pleut*). Pour cela, elle s'appuie sur des caractéristiques lexicales (p. ex. *pouvoir* est un verbe non météorologique) et/ou syntaxiques.

L'étape suivante consiste à déterminer quels sont les antécédents possibles d'un pronom référentiel. Le principe (ii) du liage nous dit qu'il ne peut se trouver dans la proposition qui contient le pronom, la recherche s'effectue donc dans les propositions ou phrases précédentes et concerne tous les syntagmes nominaux qui s'accordent en nombre et en genre avec le pronom.

En cas d'ambiguïté (plusieurs antécédents possibles), on recourt à des heuristiques basées sur la fonction grammaticale de l'antécédent (p. ex. le sujet est prioritaire par rapport à l'objet direct) et des éventuelles restrictions sélectionnelles associées à la fonction du pronom (on s'appuie sur les traits sélectionnels sémantiques des arguments du verbe, p. ex. animé, inanimé, humain, etc.). La Figure 1 montre la sortie produite par l'étiqueteur morphosyntaxique sur la séquence de phrases « *Jean a beaucoup de problèmes. Tellement qu'il ne pourra jamais les résoudre tous* ».

1	Jean	NOM-PRO-SIN-MAS	Jean	SU	
2	a	VER-IND-PRE-3-SIN	avoir		SU:Jean 1 DO:problèmes 5
3	beaucoup	DET-QUA-INN-ING	beaucoup	DO	
4	de	PRE	de		
5	problèmes	NOM-COM-PLU-MAS	problème		
6	.				
1	Tellement	ADV	tellement		
2	qu'	ADV	que		
3	il	PRO-PER-3-SIN-MAS	il	SU	Jean
4	ne	ADV	ne		
5	pourra	VER-IND-FUT-3-SIN	pouvoir		SU:il 3 SO:résoudre 8
6	jamais	ADV	jamais		
7	les	PRO-CLI-PLU-ING	le		
8	résoudre	VER-INF	résoudre		SU:il 3 DO:les résoudre les problèmes
9	tous	PRO-IND-PLU-MAS	tous		
10	.				

Figure 1 : Sortie de l'étiqueteur produite sur deux phrases, avec résolution d'anaphores

Pour chaque mot de la phrase, l'étiqueteur morphosyntaxique affiche les informations suivantes, réparties en 7 colonnes :

1. le numéro d'ordre du mot dans la phrase (utilisé pour référence),
2. la forme orthographique,
3. l'étiquette morphosyntaxique,
4. le lexème,
5. la fonction grammaticale (*SU* : sujet, *DO* : objet direct, *SO* : complément phrase, etc.)
6. les arguments (ou valence) pour les verbes, l'antécédent pour les pronoms anaphoriques,
7. la collocation identifiée (*résoudre les problèmes* dans l'exemple de la Figure 1).

L'analyseur conserve les syntagmes nominaux avec leur fonction grammaticale dans une structure de données ad hoc afin de pouvoir effectuer la RA également dans la(les) phrase(s) précédente(s). Pour une description complète de la résolution d'anaphores implémentée dans notre système, nous invitons le lecteur à consulter l'article de Wehrli, Seretan et Nerima (2020).

Références

- CHOMSKY N. (1981). *Lectures on Government and Binding*, Foris Publications.
- HOBBS J. (1978). Resolving pronoun references, *Lingua* 44:311-338. DOI : [10.1016/0024-3841\(78\)90006-2](https://doi.org/10.1016/0024-3841(78)90006-2)
- LAPPIN S., LEASS, H.J. (1994). An Algorithm for Pronominal Anaphora Resolution, *Computational Linguistics*, 20:4, 535-561. Anthologie ACL : [J94-4002](https://doi.org/10.1016/j.inffus.2020.01.010)
- MITKOV R. (2014). *Anaphora resolution*. Routledge.
- SUKTHANKER R., PORIA S., CAMBRIA E., THIRUNAVUKARASU R. (2018). Anaphora and Coreference Resolution: A Review. Consulté le 12.2.2020. DOI : [10.1016/j.inffus.2020.01.010](https://doi.org/10.1016/j.inffus.2020.01.010)
- WEHRLI E., SERETAN, V., NERIMA, L. (2020). Verbal collocations and pronominalisation. In G. C. Pastor & J-P Colson, Éd., *Computational Phraseology, IVITRA Research in Linguistics and Literature*, 20, p. 208–224. John Benjamins. DOI : [10.1075/ivitra.24.11weh](https://doi.org/10.1075/ivitra.24.11weh)

Adresse URL de la démonstration

FIPS : <http://latlapps.unige.ch/Parser/>, application « FullTagger », consulté le 24.04.2020

Spiderlex et compagnie

Sandrine Ollinger¹ Alain Polguère² Yannick Chudy³ Bruno Gaume⁴

(1) CNRS, Université de Lorraine, ATILF, UMR 7118 Nancy, F-54063, France

(2) Université de Lorraine, CNRS, ATILF, F-54063, France

(3) La Maison de l'Initiative, Toulouse, F-31100, France

(4) CNRS, Université de Toulouse, CLLE-ERSS, UMR 5263 Toulouse, F-31058, France

Sandrine.Ollinger@atilf.fr, Alain.Polguere@univ-lorraine.fr,
ychudy@gmail.com, Bruno.Gaume@univ-tlse2.fr

RÉSUMÉ

Nous présentons l'instrument de navigation Spiderlex, mis au point pour permettre une exploration dynamique des *Réseaux Lexicaux du Français et de l'Anglais* (RL-fr et RL-en). Spiderlex est à la fois un outil de visualisation des données lexicales et un instrument de travail pour les lexicographes.

ABSTRACT

Spiderlex & Co.

We introduce Spiderlex, a lexical browser designed to allow for the dynamic exploration of the French and English Lexical Networks (fr-LN and en-LN). Spiderlex is both a tool for visualizing lexical data and for the construction of Lexical Systems by lexicographers.

MOTS-CLÉS : Système Lexicaux, ressources lexicales, visualisation de réseaux, consultation.

KEYWORDS: Lexical Systems, lexical resources, network visualization, viewer interface.

1 Introduction

Les Systèmes Lexicaux (Polguère, 2014)¹ fournissent l'ensemble des informations présentes dans les modèles lexicographiques construits selon les principes théoriques et descriptifs de la Lexicologie Explicative et Combinatoire (Mel'čuk *et al.*, 1995); ils ont la particularité d'être structurés sous forme de graphes lexicaux. Ils appartiennent à la famille des réseaux lexicaux et ont les propriétés topologiques de graphes petit-monde hiérarchiques : agrégats (angl. *clusters*) et loi de puissance des degrés (Gader *et al.*, 2014). Spiderlex, développé en collaboration avec le projet Padagraph², a été conçu pour permettre une navigation interactive sur les *Réseaux Lexicaux du Français et de l'Anglais* (RL-fr et RL-en) qui prend en compte ces caractéristiques.

Il a été pensé pour répondre aux attentes variées des utilisateurs. Les lexicographes ont besoin de visualiser quotidiennement l'impact de leur travail sur la structuration générale du réseau. Les linguistes et informaticiens intéressés par l'exploitation d'un Système Lexical sont demandeurs d'un environnement leur permettant de se familiariser avec sa structure et les informations qu'il contient. Enfin, les enseignants ont besoin de ressources pédagogiques dans le cadre d'activités d'enseignement du vocabulaire ou, au niveau universitaire, pour l'enseignement de la lexicologie/lexicographie.

1. <https://lexical-systems.atilf.fr/>

2. <https://github.com/padagraph/>

Spiderlex propose donc des fonctionnalités d'accès simples, tout en permettant un paramétrage fin pour un usage avancé.

2 Présentation générale

Le RL-fr est composé à ce jour de plus de 28 500 nœuds et 77 500 arcs. Il est entièrement construit lexicographiquement, au sein du laboratoire ATILF depuis 2011. Il a déjà atteint une grande maturité, même si sa couverture du lexique français est encore partielle. Le RL-en, quant à lui, est composé de plus de 200 000 nœuds et près d'un million d'arcs. Il a été amorcé automatiquement à partir de WordNet (Fellbaum, 1998; Gader *et al.*, 2014) et fait depuis l'objet d'un travail de correction et d'enrichissement. Pour bénéficier d'une représentation graphique comme point d'entrée dans les réseaux, sans pour autant les afficher en intégralité, nous nous sommes appuyés sur les travaux de Gaume (2008) sur la proxémie pour déterminer les meilleurs sous-graphes susceptibles de présenter l'entourage d'une unité lexicale ou d'un vocable.

Pour répondre aux besoins variés des usagers, deux modes d'explorations, l'un dit *sémantique*, l'autre *lexicale*, ont été préprogrammés. L'exploration sémantique est calculée à partir des poids sémantiques (indiqués dans les métamodèles) des relations entre entités lexicales, tandis que l'exploration dite *lexicale* est calculée sans pondération. Elle tient uniquement compte de la topologie du graphe. Dans les deux modes d'exploration, chaque requête de l'utilisateur donne lieu à une *balade aléatoire* (Gaume, 2004) dans le Système Lexical. Ce parcours sillonne le graphe sans restriction vis-à-vis de l'orientation des arcs, en prenant en compte les boucles éventuelles et les arcs multiples. Par défaut, la longueur de la balade est de 3 et le nombre de nœuds retournés est de 30 par entité recherchée.

L'utilisateur de Spiderlex choisit un mode d'exploration, puis interroge les entités lexicales qui l'intéressent à l'aide de leur forme ou de leur identifiant unique. Une représentation graphique du sous-graphe obtenu est affichée dans la partie centrale de l'interface. La Figure 1 montre ainsi le résultat de la consultation du vocable TORTUE, qui regroupe deux acceptions : TORTUE I [*Que mange une jeune tortue ?*], acception qui dénote un animal et TORTUE II [*C'est une vraie tortue celui-ci, il va à deux à l'heure !*], acception métaphorique, qui dénote un individu qui se comporte avec lenteur.

Dans la partie droite de l'interface, trois onglets sont disponibles : Liens lexicaux, Agrégats et Dictionnaire. Le premier permet de sélectionner quels types de liens on veut voir apparaître dans le graphe (composant sémantique, inclusion formelle, copolysémie et fonctions lexicales³). Le sous-graphe n'est pas recalculé lorsque l'on modifie ces paramètres, mais les arcs désactivés sont rendus invisibles et les nœuds qui n'étaient reliés que par eux sont grisés. Il n'est plus possible de les sélectionner. L'utilisateur peut ainsi visualiser rapidement quelles relations structurent l'espace sémantique ou lexical qu'il est en train d'explorer.

Comme le montre la Figure 2, l'onglet Agrégats permet d'observer sous forme de listes des sous-groupes de nœuds fortement liés à l'intérieur de la sélection d'entités lexicales effectuée. Ces agrégats rendent compte de sous-groupes de nœuds fortement liés topologiquement à l'intérieur de la sélection d'entités lexicales effectuée. Ils sont calculés à l'aide de l'algorithme Walktrap (Pons & Latapy, 2006), appliqué par défaut. Tout comme les paramètres de constitution de sous-graphes, la méthode de détection de communautés est modifiable dans un mode expert, où l'algorithme Infomap (Rosvall

3. Nous vous invitons à consulter la distribution publique du RL-fr (ATILF, 2020), sa documentation et sa bibliographie pour une description détaillée des différents types de liens cités ici.

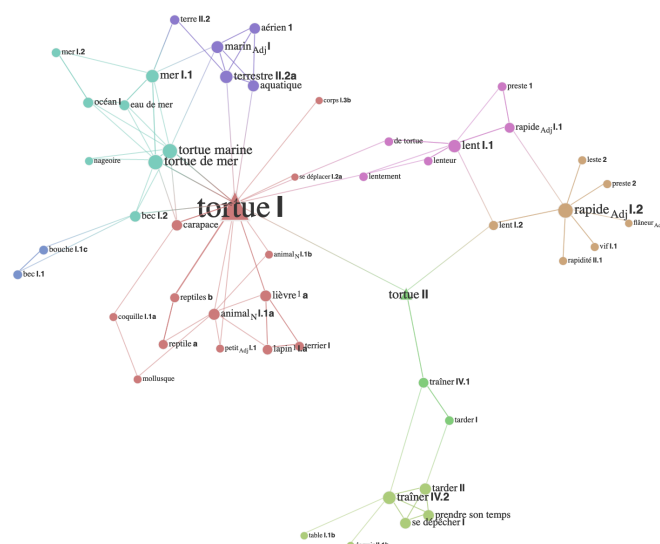


FIGURE 1 – Graphe résultant d’une exploration sémantique autour du vocable TORTUE

et al., 2009) est également disponible. On observe bien souvent que des agrégats incohérents rendent compte de descriptions insuffisantes ou de liens erronés dans la ressource.

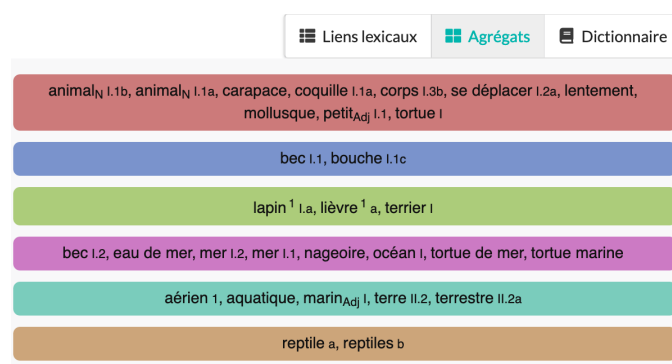


FIGURE 2 – Agrégats sémantiques dans le voisinage de TORTUE I et TORTUE II

Enfin, l’onglet *Dictionnaire*, illustré par la Figure 3 donne accès aux articles lexicographiques. Il contient l’ensemble des informations lexicales encapsulées dans chaque nœud du graphe. Dans un premier temps, les nœuds visibles sont ceux des entités lexicales recherchées. L’utilisateur peut changer d’article en sélectionnant une entité lexicale dans la vue-article ou sur la représentation graphique. Enfin, il peut lancer le calcul d’un nouveau sous-graphe en cliquant sur l’en-tête de l’article.

Spiderlex est disponible en ligne (<https://spiderlex.atilf.fr/>) et est exploité par les lexicographes des RL-fr et RL-en. Il sert aussi de source de données pour des travaux de recherche, notamment en enseignement/apprentissage du vocabulaire.

Liens lexicaux

Agrégats

Dictionnaire

tortue I

[CG]

nom commun

fém

[DF]

animal

tortue

=

animal

(petit)

dont le corps est entouré d'une carapace

qui se déplace lentement

terrestre ou aquatique

[FL]

Syn

tortue de mer, tortue marine

Contr

lièvre¹ a

Gener

animal_N 1.1b ; spéc reptiles_b

Mero

bec 1.2 ; carapace

[EX]

Puis comme elle s'était inquiétée des dommages causés aux plantes aquatiques par les

tortues d'eau qui infestaient les bassins, Ralph entreprit en fouillant à pleins bras la vase

de capturer l'une de ces bestioles.

Frantext TOURNIER Michel, *Les Météores*, 1975, p. 471

FIGURE 3 – Article lexicographique de TORTUE I

Références

- ATILF (2020). Réseau lexical du français (rl-fr). ORTOLANG (Open Resources and TOols for LANGuage) – www.ortolang.fr.
- FELLBAUM C. (1998). *WordNet : an electronic lexical database*. Language, Speech and Communication. MIT Press. DOI : [10.7551/mitpress/7287.001.0001](https://doi.org/10.7551/mitpress/7287.001.0001).
- GADER N., OLLINGER S. & POLGUÈRE A. (2014). One Lexicon, Two Structures : So What Gives ? In H. ORAV, C. FELLBAUM & P. VOSSE, Éd., *Seventh Global Wordnet Conference (GWC2014)*, p. 163–171, Tartu, Estonia : Global WordNet Association. HAL : [hal-00937187](https://hal.archives-ouvertes.fr/hal-00937187).
- GAUME B. (2004). Balades aléatoires dans les Petits Mondes Lexicaux. *Revue I3 - Information Interaction Intelligence*, **4**(3).
- GAUME B. (2008). Mapping the forms of meaning in small worlds. *International Journal of Intelligent Systems*, **23**(7), 848–862. DOI : [10.1002/int.20275](https://doi.org/10.1002/int.20275), HAL : [hal-01322013](https://hal.archives-ouvertes.fr/hal-01322013).
- MEL'ČUK I., CLAS A. & POLGUÈRE A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Paris/Louvain-la-Neuve : Duculot. DOI : [10.1017/S0008413100020582](https://doi.org/10.1017/S0008413100020582).
- POLGUÈRE A. (2014). From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography*, **27**(4), 396–418. DOI : [10.1093/ijl/ecu017](https://doi.org/10.1093/ijl/ecu017).
- PONS P. & LATAPY M. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, **10**(2), 191–218. DOI : [10.1007/11569596_31](https://doi.org/10.1007/11569596_31).
- ROSVALL M., AXELSSON D. & BERGSTROM C. T. (2009). The map equation. *The European Physical Journal Special Topics*, **178**(1), 13–23. DOI : [10.1140/epjst/e2010-01179-1](https://doi.org/10.1140/epjst/e2010-01179-1).

Utiliser les outils CORLI de conversion TEI pour l'analyse de corpus de langage oral

Parisse Christophe¹ Loïc Liégeois^{2,3}

(1) Modyco, Nanterre, France

(2) LLF, Paris, France

(3) CLILLAC-ARP, Paris, France

cparisse.parisnanterre.fr, loic.liegeois@univ-paris-diderot.fr

RÉSUMÉ

Le consortium CORLI développe des outils pour faciliter le dépôt, l'interrogation et la réutilisation des corpus oraux. Ces outils libres et open source sont basés sur la TEI comme format commun de partage. Nous présenterons deux outils différents : un outil pour la saisie et l'édition de fichiers de métadonnées et un outil permettant d'intégrer et d'utiliser des corpus de différentes sources de données transcrits dans différents logiciels.

ABSTRACT

The CORLI consortium develops tools to facilitate sharing, interrogation, and reusing of spoken language corpora. These free tools are based on TEI as a sharing format. We present two tools: a tool for metadata edition, and a tool to aggregate and use corpora transcribed in different formats and coming from different sources.

MOTS-CLÉS : TEI, PRAAT, ELAN, Transcriber, TXM, Metadonnées, Corpus.

KEYWORDS: TEI, PRAAT, ELAN, Transcriber, TXM, Metadata, Corpora.

1 Le partage de corpus de langage oral et la TEI

L'utilisation d'outils pour la recherche sur les corpus oraux est incontournable, que ce soit pour la transcription, l'édition, l'étude qualitative (PRAAT, ELAN, Transcriber, CLAN, etc.) ou pour l'analyse, soit grammaticale, soit exploratoire (outils de textométrie). Ces diverses applications à partir de formats et de conventions presque tous incompatibles requièrent des interventions manuelles lourdes et réduisent les possibilités de recherche scientifique. CORLI cherche à faciliter le travail des chercheurs en mettant à disposition des outils accessibles sans qualification informatique pour permettre de consacrer plus de temps aux analyses elles-mêmes. Les logiciels libres que nous avons développés sont tous basés sur un format commun en TEI, qui a l'avantage

d'être connu, libre, valide pour la pérennisation et assez ouvert pour s'adapter à la variété des corpus. Les outils disponibles à ce jour portent sur trois grandes fonctionnalités : l'édition des métadonnées des corpus et des enregistrements ; la conversion de formats pour l'édition des corpus ; la conversion de format pour l'exploration des corpus et l'utilisation d'outils syntaxiques sur les données convertis en TEI.

2 Trois exemples d'utilisation

L'espace de démonstration que nous proposons contiendra un poster qui présentera l'ensemble des activités de CORLI et en particulier celles qui sont liées à la problématique de l'usage de la TEI. Par ailleurs nous présenterons trois démonstrations basées sur des cas pratiques. La première concerne l'outil TEIMETA (<http://ct3.ortolang.fr/teimeta/>), les deux suivantes l'outil TEICORPO (<http://ct3.ortolang.fr/teicorpo/>). Chaque présentation permettra de préciser les compétences requises et les liens que l'on peut faire avec les sites de dépôt de corpus.

2.1 TEIMETA : création d'un fichier de métadonnées

Pour cette partie de la démonstration, nous nous mettrons dans la position d'un chercheur ayant à renseigner un ensemble de métadonnées liées à un fichier de transcription. Souvent long et fastidieux, nous verrons comment ce travail peut être facilité par l'outil au moyen par exemple de l'usage de vocabulaires contrôlés et comment le résultat s'intègre à tout fichier TEI.

2.2 TEICORPO : utilisation de PRAAT et la TEI

Il est possible de convertir des fichiers d'origine différente (PRAAT, ELAN, CLAN, Transcriber) vers la TEI et uniformisant autant que possible les transcriptions. Une analyse syntaxique basée sur TreeTagger ou sur la bibliothèque Stanford NLP (format CONLL) est également possible, avec un résultat peut alors être converti vers PRAAT pour permettre par exemple une utilisation dans PRAAT de données comprenant transcription et annotations grammaticales.

2.3 TEICORPO : utilisation de TXM et la TEI

Il est possible de rassembler par conversion des corpus issus d'origines variées pour construire une base au format TEI. Pour une utilisation scientifique pertinente, il est souvent nécessaire d'enrichir les métadonnées des corpus. En dehors de l'utilisation de TEIMETA, il est possible d'insérer des métadonnées éditées au format CSV, puis d'exporter les corpus pour leur utilisation dans TXM, avec des métadonnées requêtables et la possibilité d'écouter le son dans TXM. Un exemple basé sur le grand corpus ESLO sera présenté.

Vers un corpus optimal pour la fouille de textes : stratégie de constitution de corpus spécialisés à partir d'ISTEX

Camille de Salabert¹ Sabine Barreaux¹

(1) Inist-CNRS, 2, rue Jean Zay, CS 10310, 54519 Vandœuvre-lès-Nancy, France
camille.de-salabert@inist.fr, sabine.barreaux@inist.fr

RÉSUMÉ

Préalable indispensable à de nombreuses activités de TAL et de fouille de textes, l'élaboration d'un corpus peut nécessiter plusieurs phases de traitement pour améliorer sa qualité et ainsi obtenir les meilleurs résultats d'analyse automatique. Les post-traitements appliqués à un tel corpus, notamment pour garantir la pertinence de son contenu et l'homogénéité de son format, pourront s'avérer d'autant plus coûteux et fastidieux que la construction du corpus de travail aura été imprécise. Cette démonstration se proposera de tirer parti de la plateforme ISTEX et de ses services associés pour constituer, au travers d'un cycle itératif, un corpus homogène de publications scientifiquement pertinentes pour une utilisation simplifiée par des outils de fouille.

ABSTRACT

Towards an optimal corpus for text mining: specialized corpus building strategy from ISTEX.

The development of a corpus is an essential prerequisite for many NLP and text mining activities. This may require several processing phases to improve its quality and thus obtain better automatic analysis results. Post-processing applied to such a corpus in particular to guarantee the relevance of its content and the homogeneity of its format may be more costly and tedious if the construction of the working corpus is more basic. This demonstration will take advantage of the ISTEX platform and its associated services to build a homogeneous corpus of scientifically relevant publications for a simplified use by text mining tools through an iterative cycle.

MOTS-CLÉS : Fouille de textes, Corpus thématique, ISTEX, Affinement, Méthodologie.

KEYWORDS: Text mining, Thematic corpora, ISTEX, Refinement, Methodology.

Les algorithmes de fouille de textes promettent une accélération sans précédent de la recherche scientifique par le traitement automatique de corpus massifs de publications scientifiques qu'ils autorisent. Mais l'analyse textuelle peut se révéler compliquée quand il s'agit de traiter des corpus de grande dimension, multithématiques et multilingues, dont les textes, bien souvent non structurés et aux formats hétérogènes, sont rédigés dans des langues de spécialité. Dès lors, disposer d'un corpus tout à la fois adapté à l'objectif visé et approprié par rapport au cadre méthodologique et à l'outil considéré apparaît indispensable.

Si les guides de bonne pratique à l'intention de la communauté des linguistes s'intéressent à la construction de corpus équilibrés, représentatifs, échantillonnés ou adaptés à l'application recherchée ([Wynne, 2005](#)), peu de détails sont donnés en revanche quant à la phase de constitution

de corpus dans les travaux plus généraux de traitement automatique des langues (TAL) ou de fouille de textes et de données (TDM). La démarche mise en œuvre pour créer un corpus dépend fortement du but poursuivi et des sources utilisées, et se résume dans bien des cas à recourir à une équation basique, suivie d'étapes de nettoyage ou de reformatage, *a fortiori* dans le cas de textes intégraux (Nguyen, 2019). Dans cette démonstration, nous nous attacherons à montrer, au travers d'un cas d'usage pouvant intéresser la communauté scientifique, comment l'utilisation de la bibliothèque scientifique numérique nationale ISTE¹ et de ses services additionnels permet de créer, en adéquation avec la finalité souhaitée, un corpus raffiné au travers de cycles successifs d'exploration des résultats et d'ajustement de la requête exploitant les points forts d'ISTEX. Cette démonstration permettra *in fine* de parvenir à un corpus qui minimise les étapes de post-traitement habituellement opérées, telles que l'élimination du bruit et du silence, le reformatage des données multisources pour les rendre homogènes ou le repérage de documents au format inadéquat pour l'application ciblée.

1 Cas d'usage

À l'occasion du 250^e anniversaire de la naissance du musicien allemand Beethoven, la création d'un corpus en musicologie de publications relatives au compositeur nous offre l'opportunité d'ajouter un corpus thématique en sciences humaines et sociales aux corpus spécialisés déjà diffusés sur le site web d'exposition des données ISTE² : data.istex². L'objectif final envisagé est de valoriser ce corpus sur le web sémantique au moyen d'alignements avec les données DOREMUS. DOREMUS³ – DOing REusable MUSical data – est en effet un graphe de connaissances d'œuvres musicales interconnectées décrivant et contextualisant les catalogues musicaux de la BnF, de la Philharmonie de Paris et de Radio France qui sont ainsi partagés sur le web de données. Une étape préliminaire de reconnaissance d'entités nommées, entités propres à cet art telles que les noms de musiciens, les œuvres musicales ou les instruments de musique, ou génériques comme les dates et les lieux, sera nécessaire pour notre corpus. Elle pourrait être matière à comparer des outils de détection en fournissant un corpus d'apprentissage.

2 ISTE

Source de données pour la fouille de textes

ISTEX, archive riche de plus 23 millions de publications numériques dans toutes les disciplines et en plus de 50 langues, depuis le XIV^e siècle jusqu'à aujourd'hui, constitue une source de choix pour y rechercher les documents qui nous intéressent et leur appliquer des outils de fouille de textes, ce droit étant concédé par la licence d'utilisation de la plateforme ISTE.

Financée par l'ANR et négociée dans le cadre des licences nationales, l'initiative d'excellence en information scientifique et technique ISTE a de fait pour ambition de bâtir le socle de la bibliothèque scientifique numérique nationale et d'offrir, au travers de sa plateforme, l'accès aux collections rétrospectives de la littérature scientifique, ainsi que l'opportunité de s'en servir comme matériau de fouille à des fins de recherche scientifique.

¹ <https://www.istex.fr>

² <https://www.data.istex.fr>

³ <http://data.doremus.org>

Atouts de la ressource

Disponibles dans différents formats, fournis par les éditeurs ou produits par ISTEX, métadonnées et texte intégral des publications sont également proposés en format standard (respectivement MODS et TEI), facilitant considérablement l'exploitation de la diversité des formats reçus des 28 éditeurs présents dans l'archive.

A cette homogénéisation, sont ajoutés des enrichissements, disponibles dans un format TEI standoff, résultats du traitement de l'ensemble des collections par des outils de fouille de textes développés ou adaptés pour ISTEX ([Cuxac, 2017](#)). Ces enrichissements, de typologie plurielle, donnent notamment accès au contenu scientifique des ressources :

- catégories scientifiques attribuées respectivement à 75% et 44% des documents, soit par appariement entre un identifiant tel qu'un ISSN et une ou plusieurs catégories affectées à la publication concernée par le Web of Science, Science-Metrix ou Scopus, soit par apprentissage automatique sur les bases de données bibliographiques PASCAL et FRANCIS du CNRS ;
- entités nommées détectées dans 68% des textes en anglais ou français, grâce à une cascade de graphes mise au point pour le logiciel Unitex/CasSys ([Maurel, 2019](#)), ces entités nommées concernant les noms de personnes, lieux, organismes, indicateurs temporels, URL, etc. ;
- termes représentatifs du contenu du texte intégral, avec leur fréquence et leur spécificité, extraits de 73% des publications en anglais par TEEFT, outil d'indexation non supervisée qui dispense de la constitution problématique de ressources de référence spécialisées, nécessitées par une archive multidisciplinaire telle qu'ISTEX ;
- références bibliographiques structurées dans 58% des ressources au moyen de GROBID, outil de segmentation et de structuration fonctionnant par apprentissage automatique.

Forte de ces millions de publications numériques à valeur ajoutée, l'archive ISTEX forme une ressource majeure pour les travaux de fouilles de textes à l'usage de la communauté de l'enseignement supérieur et de la recherche.

Des services complémentaires

Toute une gamme de services a été conçue autour de l'API ISTEX afin de simplifier et de développer son utilisation, en particulier deux applications, l'une de téléchargement de corpus volumineux – ISTEX-DL –, l'autre d'exploration des corpus extraits, ainsi que d'exposition sur le web sémantique des corpus finalisés – LODEX :

- l'application ISTEX-DL⁴ – ou ISTEX Download –, interface de type formulaire intuitive et conviviale, permet facilement de télécharger jusqu'à 100 000 documents répondant à une requête, classique ou issue d'une liste d'identifiants. Le choix des formats des textes intégraux, tout comme celui des métadonnées ou des enrichissements correspondants s'opère en quelques clics, de même que la limitation du volume souhaité (comprenant l'option de sélection aléatoire d'un sous-ensemble) et la modulation du type et du niveau de compression du corpus ;
- LODEX⁵ – ou Linked Open Data ISTEX –, logiciel open source associant sémantisation et visualisation de données, est dédié quant à lui à la valorisation de données structurées ([Gregorio, 2019](#)). Cet outil transforme un corpus en site web dynamique, offrant d'une part une navigation dans ce corpus selon différents angles de vue au travers de graphiques

⁴ <https://dl.istex.fr>

⁵ <https://lodex.inist.fr>

filtrables par des facettes, et d'autre part une description et un accès aux documents du corpus. En outre, l'exposition sur le web de données autorise l'alignement avec des données similaires ou connexes, et permet le référencement et la réutilisation des corpus spécialisés.

3 Stratégie de constitution de corpus

Notre démonstration se proposera de débiter par une requête simple relative au musicien Beethoven qui ira se complexifiant, grâce à la visualisation des résultats de l'interrogation de l'API ISTEEX via son démonstrateur⁶, suivie de l'extraction à l'aide d'ISTEX-DL puis de l'analyse dans LODEX des états d'avancement successifs du corpus (états intermédiaire et affiné du corpus disponibles aux adresses suivantes : <https://beethoven-ludwigv0.corpus.istex.fr/> et <https://beethoven-ludwigv1.corpus.istex.fr/>). L'exploration à ces différentes étapes du contenu scientifique du corpus et de ses métadonnées nous permettra de proche en proche d'en déduire les correctifs à apporter à notre équation d'interrogation et nous donnera des clés pour enrichir la requête, tirant profit des qualités de la ressource ISTEEX qui assure en amont les traitements de reformatage ou de complétion des données, de différents enrichissements tels que Unitex et TEEFT afin de limiter silence comme bruit – corollaire des personnages célèbres ! – et de permettre une sélection fine de documents pertinents, et des indicateurs de qualité disponibles sur la plateforme ISTEEX au service du TDM pour ajuster aussi finement que possible le corpus à l'usage ciblé.

Ainsi, la méthodologie itérative de requêtage, d'extraction et d'exploration du corpus, permet de constituer un corpus de musicologie quasiment prêt à l'emploi pour l'extraction d'entités nommées, en réduisant le nombre des post-traitements de curation habituellement requis pour une exploitation valide.

Références

- CUXAC P. & THOUVENIN N. (2017). Archives numériques et fouille de textes : le projet ISTEEX. In *Atelier TextMine, EGC 2017 (Extraction et Gestion des Connaissances)*, Grenoble, France, January 24-27, 2017.
- GREGORIO S., COLLIGNON A., PARMENTIER F. & THOUVENIN N. (2019). LODEX : des données structurées au web sémantique. In *Atelier Web des Données, EGC 2019 (Extraction et Gestion des Connaissances)*, Metz, France, January 21-25, 2019.
- MAUREL D., MORALE E., THOUVENIN N., RINGOT P. & TURRI A. (2019). Istex: A Database of Twenty Million Scientific Papers with a Mining Tool Which Uses Named Entities. *Information*. DOI : [10.3390/info10050178](https://doi.org/10.3390/info10050178), HAL: [hal-02152978](https://hal.archives-ouvertes.fr/hal-02152978), version 1.
- NGUYEN N., GABUD R. & ANANIADOU S. (2019). COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. *Biodiversity Data Journal*, 7, e29626. DOI : [10.3897/BDJ.7.e29626](https://doi.org/10.3897/BDJ.7.e29626)
- WYNNE M., Éd. (2005). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxbow Books

⁶ <https://demo.istex.fr>

Deuxième partie

Résumés d'articles acceptés dans des conférences internationales

Analyse sémantique robuste par apprentissage antagoniste pour la généralisation de domaine

Gabriel Marzinotto^{1,2} Géraldine Damnati¹ Frédéric Béchet² Benoît Favre²

(1) Orange Labs / Lannion France

(2) Aix Marseille Univ, CNRS, LIS / Marseille France

gabriel.marzinotto@orange.com, geraldine.damnati@orange.com

frederic.bechet@lis-lab.fr, benoit.favre@lis-lab.fr

RÉSUMÉ

Nous présentons des résumés en français et en anglais de l'article ([Marzinotto et al., 2019](#)) présenté à la conférence *North American Chapter of the Association for Computational Linguistics : Human Language Technologies* en 2019.

ABSTRACT

Robust Semantic Parsing with Adversarial Learning for Domain Generalization

We present French and English abstracts of the article ([Marzinotto et al., 2019](#)) that was presented at the 2019 North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

MOTS-CLÉS : Analyse sémantique, adaptation de domaine, apprentissage antagoniste.

KEYWORDS: Semantic parsing, domain adaptation, adversarial learning.

1 Résumé en français

Cet article étudie l'amélioration de la capacité de généralisation des modèles d'analyse sémantique à travers de techniques d'apprentissage antagoniste. La création de modèles plus robustes à la variabilité inter-documents est cruciale pour l'intégration des technologies d'analyse sémantique dans les applications réelles. La question sous-jacente de cette étude est de savoir si l'apprentissage antagoniste peut être utilisé pour entraîner des modèles à un niveau d'abstraction plus élevé afin d'augmenter leur robustesse aux variations lexicales et stylistiques.

Nous proposons d'effectuer l'analyse sémantique avec une tâche adverse de classification de domaine sans connaissance explicite du domaine. La stratégie est d'abord évaluée sur un corpus français de documents encyclopédiques, annotés avec FrameNet, dans une perspective de recherche d'informations, puis sur la tâche PropBank Semantic Role Labelling sur le benchmark CoNLL-2005. Nous montrons que l'apprentissage contradictoire augmente toutes les capacités de généralisation des modèles à la fois sur les données du domaine et hors domaine.

2 English Abstract

This paper addresses the issue of generalization for Semantic Parsing in an adversarial framework. Building models that are more robust to inter-document variability is crucial for the integration of Semantic Parsing technologies in real applications. The underlying question throughout this study is whether adversarial learning can be used to train models on a higher level of abstraction in order to increase their robustness to lexical and stylistic variations.

We propose to perform Semantic Parsing with a domain classification adversarial task without explicit knowledge of the domain. The strategy is first evaluated on a French corpus of encyclopedic documents, annotated with FrameNet, in an information retrieval perspective, then on PropBank Semantic Role Labeling task on the CoNLL-2005 benchmark. We show that adversarial learning increases all models generalization capabilities both on in and out-of-domain data.

Références

MARZINOTTO G., DAMNATI G., BÉCHET F. & FAVRE B. (2019). Robust semantic parsing with adversarial learning for domain generalization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Industry Papers)*, p. 166–173, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : [10.18653/v1/N19-2021](https://doi.org/10.18653/v1/N19-2021).

