La variété stylistique en français et espagnol (corpus littéraires, analyses linguistiques automatisées et évaluation) : suite du projet

Projet répondant à l'appel d'offre

« Agorantic : Culture, Patrimoine, Sociétés Numériques » 2021

Porteur du projet : Juan-Manuel TORRES-MORENO[1]

Equipe UA : Luis-Gil MORENO-JIMÉNEZ[1], Cyrielle GARSON[4,], Graham RANGER[4], Madelena GONZALEZ[4,]

Equipe externe : Luis MENESES-LERÍN[2], Salah MEJRI[3], Lichao ZHU[3]

2 stagiaires de Master (4 mois par stage)

Laboratoires impliqués :

[1] Laboratoire Informatique d'Avignon (LIA), Avignon Université

[2] Centre de Recherche GRAMMATICA (Université d'Artois) (Arras)

[3] Membres associés GRAMMATICA

[4] Laboratoire Identité Culturelle, Textes et Théatrâlité (ICTT), Avignon Université

Email : juan-manuel.torres@univ-avignon.fr, luis-gil.moreno-jimenez@alumni.univ-avignon.fr, jluis.meneseslerin@univ-artois.fr, cyrielle.garson@univ-avignon.fr, graham.ranger@univ-avignon.fr, madelena.gonzalez@univ-avignon.fr

Objectifs : L'objectif de ce projet concerne la génération et la gestion de ressources linguistiques. En particulier nous voulons élargir des corpora en français et en espagnol existants pour étudier la variété diatopique et stylistique dans le domaine littéraire. Cette étude permettra l'identification des structures linguistiques complexes équivalentes dans les deux langues. Les corpus générés pourront être employés dans des systèmes génératifs de texte (artificiels ou naturels) ainsi que leur évaluation Nous mettrons en place un site web pour la diffusion des corpus et des outils développés.

Mots-clés : Traitement Automatique de la Langue Naturelle (TALN), Corpus littéraire, Analyse stylistique, Recherche d'Information (RI).

1. Description du Projet

Le projet est une suite au projet présenté l'année 2020. Il a comme objectif l'élargissment des corpus d'œuvres littéraires en français ainsi que le développement d'outils informatiques pour son exploitation linguistique [1]. Parmi les œuvres retenues, nous retrouvons deux cas de figure : des œuvres originales en français ou des œuvres traduites en français.

eUn premier corpus MEGALITE-fr a été constitué avec environ 1580 ouvrages et 100 millions de mots. Nous voulons élargir ce corpus pour avoir au moins 200 millions de mots et 5000 ourvrages. Pour élargir le corpus MEGALITE, une stratégie en plusieurs étapes sera continué :

1. Vérifier l'exploitabilité des ouvrages et supprimer les doublons, droits d'auteur, etc avant de les intégrer au corpus MEGALITE-fr

2. Poursuivre la récupération (semi-automatique et manuelle) de l'ensemble des documents littéraires sans aucune distinction, sous des formats qui permettent leur analyse au moyen d'instruments de calcul.

3. La deuxième étape consiste à normaliser les titres ainsi que le format des documents sous le format *utf-8* qui facilite leur traitement. Il est ainsi possible de procéder à la classification par nom d'auteur, par œuvre ou par période de publication.

4. Dans une troisième étape, poursuivre la classification de l'ensemble des documents collectés afin de diviser ceux qui ont été rédigés à l'origine en français et ceux qui ont été traduits. Cela nous permettra d'effectuer une analyse stylistique appliquée à la littérature française.

1.1 Qu'est-ce qui a été fait ?

Nous disposons actuellement du corpus **MegaLite dans 2 versions**, une version de documents littéraires en espagnol (Table 1) et une autre que le projet Agorantic de l'année passée nous a permis d'amorcer la constitution d'un corpus en Français (Table 2). Le corpus en espagnol possède une dimension adéquate (nombre de phrases, nombre de mots-type, vocabulaire étendu) qui offre la possibilité de réaliser l'analyse mentionnée ci-dessus (troisième étape). Le corpus en français commence à prendre de l'ampleur mais, les difficultés liés à l'épidémie, nous ont empêche de réaliser toutes les activités prévues.

	Phrases	Mots	Caractères
MegaLite-ES	15 M	212 M	1 262 M
Moyenne par document	3 K	41.8 K	250 K

Table 1: Corpus **MegaLite-Espagnol**, composé par 5 075 documents littéraires (M représente un valeur de 10^6 et K de 10^3)

	Phrases	Mots	Caractères
MegaLite-FR	5.2 M	100 M	621 M
Moyenne par document	3 K	45 K	150 K

Table 2: Corpus **MegaLite-Français**, composé par 1 580 documents littéraires (M représente un valeur de 10^6 et K de 10^3)

Une étude préliminaire du corpus Megalite (version avec etiquettes grammaticales) a été utilisé dans le papier : « Latent Semantic Analysis for tagging Activation States and Identifiability » qui sera publié debut 2022 dans la revue "**Journal of Intelligent & Fuzzy Systems (JIFS) Indexée dans JCR.**

Également nous avons conduit une étude pilote pour la génération de phrases littéraires crées par des personnes et par des algorithmes entraînés avec le corpus Megalite-fr (1ère version) pour évaluer :

a/ Littéracité des phrases
b/ Grammaticalité
c/ Appartenance à un contexte prédéfini
d/ Test de Turing (deviner si la phrase est artificielle ou pas)

Ainsi 300 phrases et un ensemble de 18 évaluateurs (étudiants master en informatique) sous les mêmes conditions du protocole ont été évaluées. Les résultats statistiques sont sous étude actuellement

En outre, deux papiers concernant la génération textuelle et une comparative superficielle du corpus MegaLite-Fr et MegaLite-Es (français-espagnol), ont été publiés :

Moreno-Jiménez, L. G., Torres-Moreno, J. M., Gonzalez-Gallardo Carlos-Emiliano and Wedemann, R. (2021, May). Estudio de hiperparámetros de modelos neuronales en la generación de frases literarias. In Congreso Mexicano de Inteligencia Artificial, COMIA 2021.

Moreno-Jiménez LG., Torres-Moreno JM. (2022) **MegaLite-2**: An Extended Bilingual Comparative Literary Corpus. In Arai K. (eds) Intelligent Computing. Lecture Notes in Networks and Systems, vol 283. Springer, Cham.

Manuel Alejandro Sanchez-Fernandez, Alfonso Medina-Urrea and Juan Manuel Torres-Moreno. Latent Semantic Analysis for tagging Activation States and Identifiability in Northwestern Mexican news outlets, LKE 2021 (à paraître)

1.2 Qu'est-ce qui manque à faire ?

Ce corpus littéraire permettra en plus de réaliser des analyses comparatives et contrastives pour dégager des patrons lexicaux, syntaxiques et sémantiques en tenant compte des prédicats, des arguments et des actualisateurs [2], dans l'objectif d'identifier des "moules" stylistiques [3] (Ex. : *buscar un techo* [chercher un toit], *dar el último suspiro* [mourir], etc.) [4]. Différents partenaires seront impliqués dans la mise en œuvre du projet. Le laboratoire GRAMMATICA apportera son expertise dans la phase d'analyse linguistique pour la détection de patrons stylistiques à partir du corpus. Les patrons stylistiques seront décrits à l'aide de la notion de "moule" qui permettra de croiser le lexique, la syntaxe et la sémantique afin d'étudier le style d'un auteur et/ou la variété de l'espagnol ou du français.

D'autre part, le LIA effectuera l'analyse morphosyntaxique du corpus concerné en implémentant l'étiqueteur Freeling mais aussi en effectuant une comparaison avec des étiqueteurs différents comme TreeTagger[1] afin de trouver la meilleure performance. Ce processus permettra d'identifier la catégorie grammaticale de chaque mot du vocabulaire et d'approfondir ainsi dans la détection de traits saillants du point de vue syntactico-sémantique. En outre, grâce à l'utilisation d'autres outils, il sera possible de normaliser automatiquement le format et la structure des documents du corpus afin d'optimiser et d'accélérer leur étude. Une analyse sémantique automatique est aussi prévue, cet analyse pourra se dérouler par l'implémentation de Réseaux Neuronales dédiée à cette tâche, mieux connues comme Word2vec [5].

Pour la recherche et la collecte des documents, deux stagiaires seront engagés pour une période de 3 mois chacun. Ils devront effectuer la recherche d'œuvres littéraires en français dans des banques de données publiques ou dans le cadre d'une licence permettant leur exploitation. Les documents doivent être classés en deux catégories générales : les œuvres originales en français et les œuvres traduites. Par la suite, ils devront également procéder à une classification plus détaillée au niveau du genre littéraire, en considérant les différents genres : la poésie, le roman, le théâtre, l'essai, etc.

2. Objectifs et résultats attendus

Le corpus MegaLite-ES a été utilisé dans différentes études sémantiques [6] et a été incorporé dans des travaux pour la génération de textes en espagnol [7]. Un corpus littéraire en langue française Megalite-Fr plus grand permettra de reproduire les expériences réalisées sur le corpus de l'espagnol et contribuera à la génération de textes littéraires en langue française.

- Un corpus composé de documents littéraires en français avec au moins 5000 documents littéraires.
- Étude sémantique du vocabulaire contenu dans le corpus français.
- Un ensemble de caractéristiques linguistiques (et probablement esthétiques) extraites du corpus français.
- Un ensemble de caractéristiques linguistiques extraites du corpus espagnol (MegaLite-ES)
- Une étude associative entre les caractéristiques extraites des deux corpus.
- Protocole d'évaluation de phrases littéraires crées par machines et par personnes dans un cadre élargit : 1000 phrases et une vingtaine d'évaluateurs (étudiants en lettres)

3. Caractère innovant de ce projet

Jusqu'à présent, l'étude et la constitution de corpus dans le domaine de la linguistique ont été largement abordées par la communauté scientifique [8]. Parmi ces corpus, il en existe un qui a été moins étudié que les autres, ceux composés de documents littéraires. La littérature, en raison de caractéristiques telles que la complexité du discours ou l'ambiguïté, représente un défi pour son étude ou son analyse; c'est pourquoi les corpus littéraires ne sont pas très fréquents dans les travaux de recherche, en particulier pour les langues romanes comme le français, l'espagnol et le portugais. Ce projet est donc l'occasion d'approfondir l'étude de ces ressources et, en même temps, de proposer à la communauté scientifique un ensemble de nouveaux outils/ressources qui peuvent être utilisés pour différentes tâches liées au traitement des langues [9].

Du point de vue littéraire, les textes émanant de ces outils s'inscrivent dans le champ de la littérature électronique et posent à nouveau la question pressante de l'évaluation de la

1 Outil disponible sur le site: https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/

littérature à l'ère du numérique [12,13]. A ceci s'ajoute la visée comparative du projet entre l'espagnol et le français qui permettra quant à elle d'éclairer d'autres questions d'ordre théorique et pratique, comme celles liées à la perception des genres par le lecteur ou celles ayant trait à l'idée d'un universel littéraire et artistique entre les langues [10,11].

4. Dimension interdisciplinaire

Cette étude s'inscrit dans les projets de type interdisciplinaire et cherche à faire entrer en synergie des disciplines telles que la linguistique, l'informatique, la littérature, l'espagnol et le français.

Nous pensons que le travail en équipe entre les laboratoires GRAMMATICA, ICTT et LIA permettra de mieux exploiter les ressources littéraires. D'une part, les linguistes appartenant au laboratoire GRAMMATICA fourniront toute la base de connaissances sur la langue pour effectuer une analyse critique et la détection ultérieure de caractéristiques linguistiques utiles à la communauté scientifique.

D'autre part, l'automatisation de cette analyse à partir d'une approche formelle, permettra d'intensifier les tests et de massifier les données de validation, ce qui optimisera le temps et les ressources ainsi que d'apporter un plus grand soutien aux recherches, étant donné que celles-ci ont été validées à partir d'une masse importante de données, sous une approche formelle (mathématique - computationnelle).

Le laboratoire ICTT apportera d'autres compétences d'ordre linguistique et esthétique, mais aussi en ce qui concerne l'évaluation de la partie française du corpus.

5. Positionnement dans l'Agorantic

Ce projet se positionne à l'intersection de trois axes de l'Agorantic :

- Axe 1 : Culture et numérique
- Axe 3 : Les corpus font partie du patrimoine immatériel
- Axe 5: Structuration et exploitation de corpus

Le porteur du projet est Juan-Manuel Torres, Maître de Conférences HDR en informatique au LIA. Il a une expérience dans le domaine de l'Intelligence artificielle, tout particulièrement dans le Traitement Automatique de Langues et l'Apprentissage automatique.

Une collaboration avec Cyrielle Garson Maître de Conférences en anglais, Madelena Gonzalez et Graham Ranger, au Laboratoire ICTT permettra l'évaluation du corpus français d'une part et la visée comparative avec le corpus espagnol d'autre part dans une dimension linguistique et probablement esthétique. Egalement, sera explorée l'annotation automatique des corpus au moyen des outils comme Treetagger et Freeling.

Les collaborateurs externes en Sciences Humaines, et plus particulièrement en linguistique, seront : Luis Meneses-Lerín, MCF de linguistique (FR-ES) du Laboratoire GRAMMATICA de l'Université d'Artois, Salah Mejri, Professeur en Linguistique et Lichao ZHU, chercheur postdoctoral en linguistique informatique.

Finalement, Luis Moreno-Jiménez, doctorant en informatique au LIA, participera à ce projet sur la partie concernant les méthodes employées en TAL, aussi bien pour les systèmes de génération automatique de texte (GAT) et pour les systèmes de Recherche d'information (RI).

Les stagiaires auront principalement des tâches de création, annotation et d'évaluation

pendant la phase de constitution du corpus, évaluation de phrases, ainsi que participeront aux expériences dans les articles scientifiques envisagés.

En conclusion, nous voulons :

1/ Poursuivre la construtions, l'élargissemet, la vérification et la publication du corpus MEGALITE-Fr qui a été partiellement crée

2/ Evaluer dans un protocole plus élargie la litteracité des phrases crées par des personnes et des machines

3/ Continuer à publier des résultats de nos recherches conjointes dans des congrès spécialisés

6. Budget prévisionnel

Pour bien mener ce projet nous demandons **5 000,00** euros qui seront utilisés comme suit :

1. Un financement de 1 385,00 euros pour payer pendant 4 mois un(e) étudiant(e)s Master ou Licence à Avignon Université CERI (pour l'interface web du corpus)
2. Un financement de 1 385,00 euros pour payer pendant 4 mois un(e) étudiant(e) Master à l'Université d'Artois et/ou Université d'Avignon
3. 1 500,00 euros pour le financement de prestation de services pour l'évaluation de résultats
4. 730,00 euros pour le financement de missions

7. Références

[1] Meneses-Lerín L. *Corpus et ressources numériques : nouveaux paradigmes de recherche en linguistique, en didactique et en traduction*, Studii de lingvistică, Vol. 7, Editura Universității din Oradea, 2017, 257 p.

[2] Mejri, S.: « Les trois fonctions primaires. Une approche systématique. De la congruence et de la fixité dans le langage », De la langue à l'expression : le parcours de l'expérience discursive : hommage à Marina Aragón Cobo / coord. por Cristina Carvalho, Montserrat Planelles Iváñez, Elena Sandakova; Marina Aragón Cobo (hom.), 2017, ISBN 978-84- 16724-43-7, págs. 123-144.

[3] Meneses-Lerín L. : « Les mexicanismes entre variante et langue. L'importance de la phraséologie », in Spanish Phraseology: Varieties and variations Edited by Pedro Mogorrón Huerta and Xavier Blanco, Lingvisticæ Investigationes, 38:2, 2015, pp. 331–347.

[4] ZHU L. : « Pour une notion de moule dans le figement », Édité par Giovanni Dotoli. *Les Cahiers du dictionnaire*, Classiques Garnier, n°8, 2016, p. 97-109.

[5] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Bengio, Y., LeCun, Y. (eds.) 1st International Conference on Learning Representations. ICLR, Scottsdale, Arizona, USA (2013)

[6] Moreno-Jiménez, L. G., Torres-Moreno, J. M., & Wedemann, R. (2020, June). Literary Natural Language Generation with Psychological Traits. In NLDB.

[7] Moreno Jiménez, L. G., Torres-Moreno, J. M., S. Wedemann, R., & SanJuan, E. (2020). Generación automática de frases literarias. Linguamática, 12(1), 15-30.

[8] Sierra G., Introducción a los Corpus Lingüísticos. UNAM México., 2018

[9] Aarseth, Espen, J. Cybertext: Perspectives on Ergodic Literature. Baltimore: Johns Hopkins UP, 1997.

[10] Montford, Nick. “Continuous Paper: The Early Materiality and Workings of Electronic Literature.” Modern Language Association Conference. 2004. http://nickm.com/writing/essays/continuous_paper_mla.html

[11] Simanowski, Roberto. "Hellopoetry, Bio Poetry and Digital Literature: Close Reading and Terminological Debates." The Aesthetics of Net Literature: Writing, Reading and Playing in Programmable Media. Ed. Peter Gendolla and Jergen Schafer. Bielefeld: Transcript, 2007. 43-66.

[12] Stalybrass, Peter, et al. Language Machines: Technologies of Literary and Cultural Production. New York: Routledge, 1997.

[13] Wardrip-Fruin, Noah. “Digital Media Archaeology: Interpreting Computational Processes”. Media Archaeology. Ed. Erkki Huhtamo and Jussi Parikka. Berkeley : University of California Press, 2011. 302–322.

[14] Manuel Alejandro Sanchez-Fernandez, Alfonso Medina-Urrea and Juan Manuel Torres-Moreno. Latent Semantic Analysis for tagging Activation States and Identifiability in Northwestern Mexican news outlets, LKE 2021 (à paraître)

[15] Moreno-Jiménez, L. G., Torres-Moreno, J. M., Gonzalez-Gallardo Carlos-Emiliano and Wedemann, R. (2021, May). Estudio de hiperparámetros de modelos neuronales en la generación de frases literarias. In Congreso Mexicano de Inteligencia Artificial, COMIA 2021.

[16] Moreno-Jiménez LG., Torres-Moreno JM. (2022) **MegaLite-2**: An Extended Bilingual Comparative Literary Corpus. In Arai K. (eds) Intelligent Computing. Lecture Notes in Networks and Systems, vol 283. Springer, Cham.