TAL et humanités numériques

Jean-Gabriel Ganascia* — **Francesca Frontini** **

* *LIP6 - Sorbonne Université et CNRS*
Jean-Gabriel.Ganascia@lip6.fr
** *Laboratoire Praxiling - Université Paul Valéry Montpellier 3 et CNRS*
francesca.frontini@univ-montp3.fr

1. Introduction

C'est un peu avant le tournant du millénaire, dans la fin des années quatre-vingt-dix, que le terme « humanités numériques » (HN) (*digital humanities* en anglais) fait son apparition de façon massive. Il évoque le virage moderniste et informatique des humanités consécutif à la numérisation des contenus, le terme « humanités » étant entendu au sens anglo-saxon d'étude des œuvres humaines. Pour autant, l'utilisation de l'informatique dans les disciplines relevant des humanités est bien antérieure. Elle remonte à la fin des années quarante, avec le projet *Index Thomisticus* du père Roberto Busa qui se proposait, dès 1949, d'indexer la *Somme théologique* de Thomas d'Aquin à l'aide d'ordinateurs. Quant à l'utilisation de statistiques et de nombres pour étudier les textes littéraires, elle est plus ancienne encore. Ainsi, évoque-t-on parfois les travaux d'Augustus de Morgan qui proposa, dès 1851, une étude quantitative de la fréquence des mots pour caractériser le style des auteurs. En somme, cela fait longtemps que les humanités sont numériques, et ce, dans le double sens du terme, à la fois parce qu'elles emploient des nombres et parce qu'elles recourent aux technologies de l'information et de la communication.

Initialement, ces travaux ont porté sur le texte à la fois par commodité, parce que ce sont les contenus les plus faciles à numériser, et par habitude, parce que cela reconduisait les anciennes pratiques de la philologie, avec la construction d'index, et de la linguistique, avec l'établissement de lexiques. Le champ de l'« informatique littéraire et linguistique » (*Literary and Linguistic Computing* en anglais) (Hockey, 2004) résume bien, dans son intitulé, ce croisement entre le calcul, les disciplines d'érudition, dont les études littéraires font partie, et les sciences du langage. Toutefois, avec le

TAL. Volume 60 – n°3/2019, pages 7 à 12

temps, ce domaine a subi de multiples évolutions. D'un côté, les études sur la langue prirent leur autonomie et s'agrégèrent aux efforts très précoces d'automatisation de la traduction, ce qui donna naissance au traitement automatique des langues (TAL), d'un autre côté les travaux dans les HN s'étendirent à d'autres contenus, en particulier à des contenus multimodaux, images bi et tridimensionnelles, vidéos, sons, etc.

Depuis une vingtaine d'années, le champ des HN est en rapide expansion et ses frontières sont à la fois difficiles à identifier et en constante évolution (Dacos et Mounier, 2015 ; Terras *et al.*, 2013 ; Ganascia, 2015). Du fait de la numérisation des contenus et de la possibilité de les traiter avec des ordinateurs, les humanités se transforment, en particulier les études littéraires, l'histoire, l'archéologie, la sociologie, et cela ouvre la voie à l'émergence de nouvelles pratiques scientifiques que l'on range sous le vocable d'humanités numériques.

Dans ce contexte, même si un certain nombre d'œuvres humaines, qu'il s'agisse de tableaux, d'objets, par exemple de poteries, nous sont données sous forme multimodale, la plupart d'entre elles, que ce soit en littérature, en philosophie, en archéologie ou en histoire, nous parviennent sous forme textuelle. De ce fait, les techniques de traitement des textes sont essentielles pour les HN. Et, parmi ces techniques beaucoup recourent aux techniques du TAL, qui paraissent dès lors d'un immense profit pour les HN et en particulier pour le très vaste sous-domaine des HN que l'on qualifie d'« humanités numériques textuelles ».

Cependant, alors que la recherche actuelle en TAL s'articule généralement autour de tâches bien identifiées et plus ou moins complexes (comme la reconnaissance d'entités nommées, l'analyse syntaxique, l'extraction d'informations, les systèmes questions-réponses, le résumé de texte, etc.), les HN utilisent des techniques et des méthodes de TAL comme outils, en les hybridant à d'autres techniques issues de la fouille de données, de l'algorithmique des chaînes de caractères ou de la théorie des graphes, et en les intégrant dans des scénarios de recherche complexes, allant de l'acquisition à l'annotation et à l'analyse de textes, ces derniers incluant aussi bien des collections de textes bruts, que des éditions numériques hautement encodées. En conséquence, les défis ultimes des HN ne visent pas uniquement à améliorer les performances des outils de TAL, mais aussi, et surtout, leur utilisation dans les différents champs des humanités. Au-delà, la taille des corpus varie considérablement, depuis de grandes bibliothèques comprenant des centaines de milliers d'ouvrages numérisés – avec malheureusement de trop fréquentes erreurs – à de petits ensembles de dizaines ou de centaines de livres. Citons, à titre d'illustration, les travaux d'attribution d'auteurs appliqués aux manuscrits médiévaux de (Pinche *et al.*, 2019), de stylistique outillée sur la poésie espagnole (Ruiz *et al.*, 2017), ou le développement d'études de narratologie dans les romans fondées sur l'emploi de techniques de détection d'entités nommées (voir entre autres (de Does *et al.*, 2017) et (Alex *et al.*, 2019)).

À ces différences de finalité, s'ajoutent la très grande variété et complexité des textes traités. La diversité des types de textes communément traités par les HN, diversité d'époques, de registres ou de genres (poésie, théâtre, etc.), constitue souvent, par sa nature, un défi supplémentaire pour les outils et algorithmes courants. En particu-

lier, les documents historiques consignés dans des variantes linguistiques anciennes peuvent poser des problèmes tant d'un point de vue linguistique que pour la complexité de leur contenu. Et, il en va de même avec les textes littéraires, en particulier avec la poésie, du fait des contraintes métriques et des licences grammaticales qu'elle autorise. Il s'ensuit que des opérations désormais assez bien maîtrisées dans le champ du TAL, comme l'étiquetage syntaxique, la lemmatisation ou la racinisation (*stemming*), présentent, dans le contexte des HN, de nouveaux défis, lorsque les corpus annotés se font rares.

Enfin, malgré, ou plutôt du fait de toutes ces difficultés, les applications des HN peuvent se présenter elles-mêmes comme un banc d'essai idéal pour évaluer les dernières avancées dans le TAL. Cela paraît crucial aujourd'hui, à l'heure où la reproductibilité des résultats se pose avec acuité en sciences en général, et en TAL en particulier (Kovár *et al.*, 2016 ; Cohen *et al.*, 2018). En effet, du fait de la variété des corpus et de la profonde connaissance que les spécialistes des disciplines d'érudition ont de leurs propres corpus, on peut tester efficacement des techniques de TAL dans ces domaines et s'interroger sur la pertinence des méthodologies que l'on met en œuvre.

2. Présentation des articles

Ce numéro spécial de la revue TAL présente une petite anthologie des recherches situées à la croisée des chemins entre les HN et le TAL, un accent particulier étant mis sur des projets dans lesquels les outils du TAL sont développés et/ou appliqués pour annoter, traiter et étudier des contenus textuels provenant de différentes disciplines des humanités. Le parcours que nous proposons vise à mettre en évidence l'apport du TAL en montrant qu'il peut servir à maints égards dans le champ des HN[1] – cela va de la transcription automatique à l'annotation, à l'exploration, à l'analyse sémantique, à l'extraction et à la modélisation de connaissances. Nous verrons aussi que le TAL sert de support à des approches très variées des HN, où l'on aborde toutes sortes de genres littéraires, depuis la poésie homérique, jusqu'aux traductions et aux textes géographiques, écrits dans de multiples variétés linguistiques depuis les textes médiévaux, écrits en latin ou en langue vernaculaire, jusqu'au français moderne. Si la sélection n'a évidemment pas de présomption d'exhaustivité, ne pouvant pas représenter toutes les tendances actuelles, elle offrira sans doute au lecteur une idée de l'ampleur des recherches qui sont en train de définir ce secteur.

Transcription automatique et segmentation thématique de livres d'heures manuscrits (Daille et al.*)*

Ce premier article présente un exemple paradigmatique en ce qu'il montre comment une équipe de recherche, composée de philologues computationnels et de ta-

1. Pour un approfondissement voir le recensement systématique des « *Research Activities* » dans la taxonomie TaDiRAH - *Taxonomy of Digital Research Activities in the Humanities* - `http://tadirah.dariah.eu/vocab/index.php`

listes, s'attaque à la première étape de la chaîne d'extraction d'information textuelle à partir de documents numérisés, à savoir à la segmentation du texte. Il s'agit de procéder à l'analyse de la mise en page et à la reconnaissance de l'écriture dans les manuscrits des livres d'heures, « plus grand best-seller de tout le Moyen Âge », pour citer nos auteurs. Ce travail recourt à de nombreux algorithmes, depuis les réseaux de neurones profonds jusqu'aux chaînes de Markov cachées, et à des approches semi-supervisées. Dans tous les cas, la connaissance approfondie des textes et de leur mise en page facilite la mise en œuvre des algorithmes et permet de choisir la solution la plus appropriée.

Édition comparative intermédiale de séries traductives : exploiter les homologies pour créer des visualisations modulables (Suchecka et al.*)*

La numérisation des textes et l'identification de leur structure sont évidemment préalables à toute analyse outillée des textes. Dans cet article les auteurs nous proposent de comparer différentes traductions françaises du dixième livre des *Métamorphoses* d'Ovide à l'aide de deux outils d'alignement textuel ; ces traductions, équivalentes quant au contenu, ou supposées telles, mais différentes du point de vue linguistique et lexical, permettent de tester les algorithmes d'alignement. L'analyse montre aussi comment les outils TAL proposés aux humanistes doivent tenir compte de contraintes pratiques propres aux HN, notamment de l'encodage en TEI qui se superpose au contenu textuel pur pour y ajouter des éléments structuraux de mise en page et d'édition, dont les systèmes de comparaison devraient pouvoir tirer profit.

Vector space models of Ancient Greek word meaning, and a case study on Homer (Rodda et al.*)*

En poursuivant notre parcours ascendant vers des niveaux d'analyse linguistique de plus en plus élevés, nous arrivons à cette contribution de philologie numérique en langue anglaise. Il s'agit là d'utiliser la sémantique distributionnelle pour explorer des aspects de type lexical, notamment liés à la phraséologie de la langue grecque homérique et en particulier à la récurrence de formules plus ou moins figées comme « Achille aux pieds légers » et aux structures annulaires que leur répétition produit. Dans l'expérience proposée, différents modèles distributionnels sont confrontés à une référence « idéale » dérivée du travail des lexicographes anciens et modernes, afin de pouvoir identifier le paramétrage optimal pour cette tâche spécifique. Si la connaissance philologique nécessaire pour adapter les algorithmes est considérable, l'approche quantitative montre néanmoins ses avantages, car elle permet d'évaluer systématiquement différents aspects - tels la flexibilité sémantique des expressions – qui n'avaient jamais été pris en considération jusque là.

Chronique d'un échec : identification des métaphores dans les écrits des géographes (Mpouli)

Si les articles présentés jusqu'ici offrent une perspective plutôt positive des apports du TAL aux HN, il est important de souligner aussi les difficultés, en particulier

celles qui tiennent à des spécificités textuelles. Cet article explore la délicate question de la détection et de l'annotation des métaphores – et plus généralement de toutes les figures tropes – qui restent l'un des défis les plus complexes en TAL. L'approche choisie se fonde sur l'allocation de Dirichlet latente et vise à identifier les contextes métaphoriques dans lesquels un écart sémantique entre le domaine cible et le domaine source est identifiable. Cette méthode, très utilisée en extraction d'informations, ne semble pas toutefois donner les résultats attendus à cause de la haute spécificité des typologies et sous-typologies textuelles dans le domaine de la géographie, ce qui rend difficile l'apprentissage automatique des domaines alignés. La conclusion est que la transposition immédiate aux HN d'algorithmes et de solutions existantes n'est pas toujours possible et que seule une analyse approfondie permet de circonscrire la question de recherche afin de proposer des solutions plus adéquates.

The names of lighting artefacts : extraction and representation of Portuguese and Spanish terms in the archaeology of al-Andalus (Almeida et al.*)*

Le dernier des cinq articles proposés dans ce numéro spécial nous présente une tout autre perspective, qui croise la linguistique de corpus, l'utilisation de ressources linguistiques numériques, la modélisation des connaissances et les humanités numériques. L'article propose un schéma qui est désormais de plus en plus typique dans les HN : un corpus de textes anciens est constitué puis traité à l'aide d'algorithmes et d'outils d'extraction lexicale avant qu'une terminologie du domaine en soit extraite et que celle-ci soit ensuite modélisée par les experts à l'aide d'une ontologie formelle. Au-delà du sujet fascinant sur lequel porte cet article, à savoir les luminaires dans la culture andalouse, ce travail est représentatif d'une approche de la modélisation qui établit un lien entre l'information linguistique et l'information conceptuelle, tout en préservant la distinction entre les deux plans, celui de la langue et celui des concepts. Cette approche, qui est largement utilisée tant dans les HN qu'en linguistique informatique, est à l'origine des schémas du modèle *Ontolex-Lemon*, qui est maintenant très utilisé pour la représentation des ressources lexicales et leur exploitation avec le TAL.

3. Remerciements

Nous remercions le comité éditorial et scientifique de la revue TAL, ainsi que le comité scientifique invité, en particulier les relecteurs, qui ont contribué par leur temps et leurs efforts à la qualité de ce numéro.

Comité de lecture : Adrien Barbaresi (Berlin-Brandenburg Academy of Sciences), Valérie Beaudouin (Télécom ParisTech), Federico Boschetti (Istituto di Linguistica Computazionale « A. Zampolli » CNR, Pisa), Sascha Diwersy (Université Paul-Valéry Montpellier 3), Antoine Doucet (Université de La Rochelle), Maud Ehrmann (École polytechnique fédérale de Lausanne), Clovis Gladstone (University of Chicago), Agata Jackiewicz (Université Paul-Valéry Montpellier 3), Adam Jatowt (Kyoto University), Mike Kestemont (University of Antwerp), Anas Fahad Khan (Istituto di

4. Bibliographie

Alex B., Grover C., Tobin R., Oberlander J., « Geoparsing Historical and Contemporary Literary Text Set in the City of Edinburgh », *Language Resources and Evaluation*, vol. 53, n° 4, p. 651-675, December, 2019.

Cohen K. B., Xia J., Zweigenbaum P., Callahan T. J., Hargraves O., Goss F., Ide N., Névéol A., Grouin C., Hunter L. E., « Three Dimensions of Reproducibility in Natural Language Processing », *Proceedings of the International Conference on Language Resources & Evaluation (LREC 2018)*, vol. 2018, p. 156-165, May, 2018.

Dacos M., Mounier P., Humanités Numériques : État des lieux et positionnement de la recherche française dans le contexte international., Research Report, Institut français, March, 2015.

de Does J., Depuydt K., van Dalen-Oskam K., Marx M., « Namescape : Named Entity Recognition from a Literary Perspective », *in* J. Odijk, A. van Hessen (eds), *CLARIN in the Low Countries*, Ubiquity Press, p. 361-370, 2017.

Ganascia J.-G., « The Logic of the Big Data Turn in Digital Literary Studies », *Frontiers in Digital Humanities*, vol. 2, p. 7, 2015.

Hockey S., « The History of Humanities Computing », *in* S. Schreibman, R. Siemens, J. Unsworth (eds), *A Companion to Digital Humanities*, Blackwell, Oxford, 2004.

Kovár V., Jakubícek M., Horak A., « On Evaluation of Natural Language Processing Tasks - Is Gold Standard Evaluation Methodology a Good Solution ? », *Proceedings of the 8^{th} International Conference on Agents and Artificial Intelligence (ICAART 2016)*, p. 540-545, 2016.

Pinche A., Camps J.-B., Clérice T., « Stylometry for Noisy Medieval Data : Evaluating Paul Meyer's Hagiographic Hypothesis », *Digital Humanities Conference 2019 - DH2019*, ADHO and Utrecht University, Utrecht, Netherlands, July, 2019.

Ruiz P., Martínez Cantón C., Poibeau T., González-Blanco E., « Enjambment Detection in a Large Diachronic Corpus of Spanish Sonnets », *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Association for Computational Linguistics, Vancouver, Canada, p. 27-32, August, 2017.

Terras M., Vanhoutte E., Nyhan J., *Defining Digital Humanities : A Reader*, Routledge, London/New York, 2013.