Digital Studies / Le champ numérique

Reboul, Marianne. 2024. "Alignement sémantique et manque de données : l'apport des modèles de langue. Le cas du latin et du grec," in "DH Unbound 2022, Selected Papers," ed. Barbara Bordalejo, Roopika Risam, and Emmanuel Château-Dutier, special issue. *Digital Studies/Le champ numérique* 13(3): 1–13. https://doi.org/10.16995/dscn.9669.

Open Library of Humanities

Alignement sémantique et manque de données : l'apport des modèles de langue. Le cas du latin et du grec

Marianne Reboul, IHRIM, École Normale Supérieure de Lyon, FR, marianne.reboul@ens-lyon.fr

Le présent article vise à proposer aux spécialistes en sciences humaines et sociales, en particulier à ceux qui traitent de corpus disposant de peu de données, des méthodes récemment développées en apprentissage machine, spécifiquement pour les besoins des sciences humaines. Nous nous attachons spécifiquement à la création d'espaces sémantiques vectoriels pour les langues anciennes.

This article aims to provide humanists and social scientists, particularly those dealing with low-resource corpora, with recently developed machine learning methods specifically for the humanities. We focus specifically on the creation of vector semantic spaces for ancient languages.

Introduction

Le présent article vise à proposer aux spécialistes en sciences humaines et sociales, en particulier à ceux qui traitent de corpus disposant de peu de données, des méthodes récemment développées en apprentissage machine pour mesurer la proximité sémantique entre plusieurs langues (voire l'évolution sémantique).

Pour cet article, nous proposons d'identifier des similarités et différences sémantiques entre des textes philosophiques grecs et latins de l'époque classique (pour le grec, du Ve siècle av. J.-C. au IIIe siècle av. J.-C., et pour le latin, du IIIe siècle av. J.-C. au Ier siècle ap. J.-C.). De cette manière, nous pouvons visualiser l'évolution de certains thèmes dans le temps entre les deux langues. Il s'agit d'un corpus homogène quant aux thèmes abordés, sans surreprésentation d'un auteur sur un autre, avec un spectre chronologique large. Cet article a aussi pour vocation d'expliquer quelles sont les méthodes qui peuvent être employées pour visualiser certaines notions clés alors que l'on ne dispose que d'un micro-corpus. Voici une représentation chronologique des auteurs que nous étudions (voir **Figure 1**), tous disponibles sur Perseus (Crane 1987).

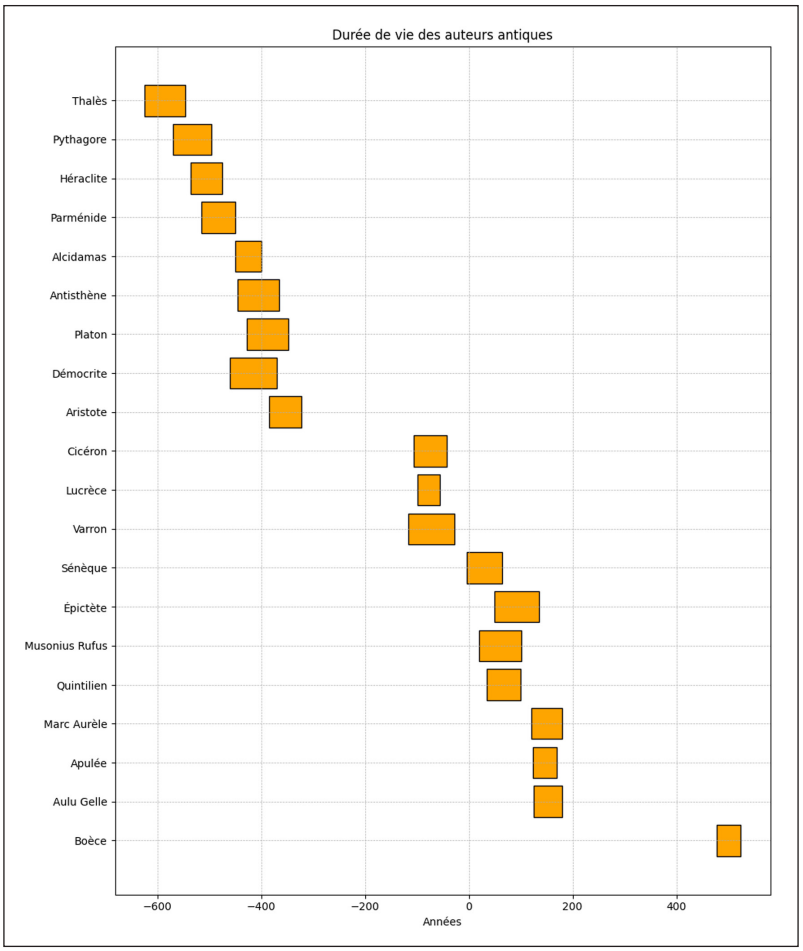

Figure 1 : Diagramme de Gantt représentant les auteurs grecs et latins étudiés

Les quantités de données disponibles sont maintenant telles qu'elles sont impossibles à analyser par le seul intermédiaire humain. Pour les lettres classiques, entre autres, de très nombreux outils sont actuellement développés dans de nombreux langages de programmation, notamment avec Python et R. Nous pensons, par exemple, au lemmatiseur et POStagger développé par Enrique Manjavacas (Manjavacas, Kádár, et Kestemont 2019), ou à son intégration pour le latin et le grec par Thibault Clérice (Clérice et Camps 2022). Par exemple, pour le simple cas de l'analyse du grec ancien ou du latin, de nombreux modules efficaces existent entre autres en Python, à commencer par les outils d'annotation comme CLTK (Johnson et al. 2021) ou stanza (Qi et al. 2020). De même que pour la pérennisation et la publication des textes existent de façon de plus en plus répandue des normes telles que la norme EpiDoc TEI (Elliott, Bodard, et Cayless 2023). Il s'agit là pour l'essentiel d'outils d'analyse automatique de la syntaxe, voire d'analyse sémantique dans certains cas (Name Entity Recognition et plongements de mots de CLTK), mais ces outils ne permettent pas, en tant que tels, d'analyser automatiquement des corpus multilingues. C'est sur ce point que cet article apporte sa contribution.

Dans cet article, nous recourons à deux méthodes d'alignement des vecteurs de mots, d'abord un alignement par combinaisons d'espaces sémantiques monolingues (avec MUSE [Conneau et al. 2017]), puis un alignement grâce à un modèle massif multilingue BERT (Devlin et al. 2019) que nous affinons à partir de données parallèles (pour l'instant non publiables). Nous prétraitons tous les textes avec stanza. Nous passons, pour la première méthode, des listes de phrases lemmatisées, et pour la seconde méthode des listes de phrases non-lemmatisées (car BERT est plus efficace pour représenter des mots en contexte lorsque les flexions sont maintenues) que nous corrélons aux équivalents lemmatisés (nous regroupons les tokens non-lemmatisés en les moyennant sous une seule étiquette avec le lemme correspondant, en conservant leurs indices). Nous entraînons ensuite les vecteurs globalement d'abord, puis par tranches temporelles (nous scindons le corpus en deux, de sorte à obtenir une cohérence sur le plan temporel ainsi que sur la masse de données prise en compte). Nous ne pouvons pas publier le modèle obtenu pour l'instant, mais nous publions l'intégralité du code, ainsi que les fichiers de sauvegarde des embeddings obtenus, sur un dépôt GitHub (Reboul 2023). Le code tel qu'il est soumis permet deux types de visualisation des résultats, un en script, et un, grâce à l'export des données pour un visualiseur d'espaces vectoriels tel que celui de tensorflow, nommé « projector » (Abadi et al. 2015).

Les plongements de mots, non-contextuels et contextuels

L'ensembles des techniques qui seront présentées dans cet article n'ont pas été développées spécifiquement pour l'analyse des textes anciens, et leur usage doit donc être adapté à leur objet.

Nous nous proposons d'étudier l'évolution du sens des mots dans les langues anciennes indo-européennes. Pour ce faire, nous partons de la théorie distributionnelle (Firth 1957) de la langue, qui pose que le sens d'un mot dépend du contexte qui l'entoure. Autrement dit, si nous connaissons le sens des mots qui constituent le contexte d'un mot inconnu, nous sommes capables, par inférence, de déduire le sens du mot inconnu, ou du moins d'en avoir une compréhension approximative. Une des techniques qui permet de représenter le sens d'un mot en fonction de son contexte est la représentation de « plongements de mots », parfois plus connue sous le nom de « word embeddings ».

L'idée, simplifiée, est que, si nous nous figurons que le langage est un espace géométrique où chaque mot a une place en fonction de son emploi, nous pouvons déterminer la distance sémantique qui existe entre les différents mots. Si nous donnons à chaque mot une représentation numérique, nous pouvons le transformer en vecteur (c'est-à-dire une direction dans un espace en plusieurs dimensions). Les valeurs constituant le vecteur de chaque mot sont obtenues en fonction de la présence du mot dans un certain contexte. Conséquemment, des mots ayant une représentation distributionnelle proche (c'est-à-dire apparaissant dans un même contexte) doivent être des vecteurs similaires, et donc avoir des sens similaires. Par exemple, nous pouvons supposer que, si nous prenons les vecteurs « Jupiter », « Zeus », « Athéna » et « Minerve » dans un corpus français, les vecteurs de « Jupiter » et de « Zeus » seront plus proches dans l'espace sémantique que les vecteurs d'« Athéna » et de « Minerve », qui ne seront pas employés dans un contexte comparable.

Il existe différents outils permettant de créer des plongements de mots, ou « word embeddings », Word2Vec (Mikolov et al. 2013) et GloVe (Pennington, Socher, et Manning 2014) étant parmi les plus connus. Dans cette expérience nous utilisons les vecteurs FastText (Bojanowski et al. 2016). Ils procèdent de manière différente, mais obéissent à un même principe : la représentation mathématique d'un mot s'effectue en fonction de son contexte, et une fois la représentation vectorielle de chaque mot effectuée dans l'espace sémantique, il est possible de réaliser des opérations mathématiques sur chacun des vecteurs. Par exemple, il est possible de mettre en lumière des relations analogiques entre les vecteurs, l'exemple le plus connu étant le suivant : si nous avons la représentation vectorielle de « femme » et d'« homme », ainsi que celle de « roi », nous pouvons calquer la relation entre le vecteur « femme » et le vecteur « homme » sur le vecteur « roi », en soustrayant le vecteur « homme » au vecteur « roi » et en y ajoutant le vecteur « femme ». Le résultat attendu serait alors le vecteur « reine ». Nous pourrions en déduire que « roi » est à « homme » ce que « reine » est à « femme ».

Dans l'exemple qui précède, nous avons évoqué un certain type d'« embedding », que l'on qualifie de « statique ». Cela signifie qu'à chaque notion correspond un

« embedding » particulier. Autrement dit, un « embedding » selon ce mode de calcul n'a qu'une seule et unique représentation vectorielle. Il existe un autre type d'« embedding », très utilisé actuellement, qui est l'« embedding » contextuel. Les « embeddings » contextuels sont encapsulés dans une séquence entière de mots-contexte dont ils sont indissociables. Ainsi l'« embedding » d'un mot particulier ne sera pas le même tout au long du corpus : un mot pourra avoir autant d'« embeddings » que d'occurrences. Cette méthode permet une compréhension plus fine de l'usage d'un mot en contexte, et est notamment très efficace pour la classification textuelle ou l'association sémantique de séquences (Miaschi et Dell'Orletta 2020). L'ensemble des vecteurs contextuels obtenus pour un corpus peut ainsi former un modèle de langue contextuel (comme avec les modèles massifs comme BERT ou les modèle XLM-R [Conneau et al. 2020]). Bien que les modèles de type BERT ne soient pas spécifiquement prévus pour obtenir des vecteurs de mots, il est tout de même possible d'obtenir des représentations globales d'un seul mot. Dans notre cas, il est nécessaire d'obtenir un seul et même vecteur par entité, ce qui est théoriquement contrintuitif avec l'utilisation de vecteurs contextuels.

Pour obtenir des vecteurs fixes et unis sous une même entité en reprenant les modèles massifs de BERT (que nous affinons ensuite sur nos propres corpus), nous proposons de concaténer les états cachés (les « hidden states », les sorties de chaque couche du « transformer ») lors de l'entraînement de chaque vecteur, et regroupons les vecteurs avec entrée similaire avec une opération de « clustering » de type K-means. Une telle approche a trois avantages : d'abord, elle nous permet de regrouper les vecteurs contextuels d'un mot sous une seule et même étiquette, puis elle conserve *a minima* la finesse de variation contextuelle obtenue à chaque « époque » de l'analyse (les stades de l'entraînement des vecteurs), et enfin elle nous permet d'avoir recours aux modèles massifs multilingues proposés en libre accès.

Espace sémantique et multilinguisme

La grande difficulté de ce type d'expérience est l'aspect multilingue : il est relativement aisé d'obtenir des espaces vectoriels monolingues à l'heure actuelle sur des corpus de taille variable. En revanche, il est nettement plus complexe, notamment lorsque les corpus sont de nature et de taille différente, d'obtenir des espaces sémantiques multilingues. En effet, l'obtention d'espaces vectoriels multilingues posent plusieurs problèmes. D'abord, sur le plan théorique, même si les vecteurs représentés sur un hypothétique espace multilingue sont mathématiquement comparables, il est nécessaire d'avoir une excellente connaissance du corpus d'entraînement, pour être sûr que les objets comparés soient effectivement sémantiquement comparables. Par

exemple, si un espace est créé à partir de corpus restreints, les biais de corpus seront d'autant plus forts que les données seront moindres. Comparer des vecteurs latins et grec formés à partir de corpus très différents et restreints n'aurait pas de sens sur le plan épistémologique. Un autre problème est celui de la fiabilité des résultats lorsque les données d'entraînement des modèles sont peu nombreuses. Pour entraîner un modèle multilingue, il faut disposer de données parallèles suffisantes pour orienter les décisions du moteur. En l'absence de données parallèles nombreuses, les approches traditionnelles peuvent donner de médiocres résultats.

Nous proposons deux méthodes différentes pour ce cas d'étude. La première option est d'aligner deux espaces vectoriels monolingues l'un sur l'autre. C'est ce que propose notamment Alexis Conneau avec MUSE (Conneau et al. 2017). Le principe qui sous-tend cette option est que les langues fonctionnent sémantiquement de la même manière, et qu'une même réalité tend à s'exprimer, distributionnellement parlant, de la même façon entre les différentes langues. Dans cette optique, les langues sont considérées comme isomorphes, et donc, potentiellement, alignables. Bien entendu, cette assertion ne peut être soutenue que dans la mesure où les espaces sémantiques à aligner sont sémantiquement comparables, comme nous l'avons mentionné précédemment.

Il s'agit donc, dans cette optique, d'entraîner deux modèles distincts, un pour le latin, un pour le grec, sur un corpus restreint mais comparable sémantiquement, à savoir un corpus de textes philosophiques. Nous obtenons donc un espace sémantique unique pour chaque langue. Pour superposer les espaces, nous effectuons une rotation de l'espace cible (en changeant son orientation sans altérer les distances entre les vecteurs) afin qu'il se superpose à l'espace source. Deux étapes sont nécessaires. La première étape est celle de l'« adversarial training », qui est une forme d'apprentissage supervisé : l'on donne à notre modèle des contrexemples, pour inviter le modèle à effectuer une discrimination et à apprendre de ses éventuelles erreurs. L'adversaire (la création de contrexemples) essaie de maximiser les erreurs en transformant les données, avec certaines conditions sur la manière dont il les transforme (par exemple, l'adversaire peut faire des rotations dans le plan de l'espace sémantique que le discriminateur utilise pour faire ses prédictions, conservant ainsi les distances entre les vecteurs mais pas les directions). Le modèle devient donc de plus en plus robuste et ajuste la rotation d'un espace sur l'autre. La seconde étape est l'analyse de Procruste, qui consiste alors à comparer deux formes, déformant un objet pour le rendre autant que possible semblable à sa cible (ici en se basant sur la fréquence relative de certains termes, qui agiront comme points d'ancrage). Enfin, nous amplifions les distances entre les vecteurs lorsque ceux-ci sont situés dans un espace à forte densité, pour les rendre plus spécifiques.

Cette option est utile lorsque les données parallèles d'entraînement sont très peu nombreuses. En effet, l'implémentation proposée notamment par le module MUSE peut être entièrement non-supervisée ou partiellement supervisée avec très peu d'entrées parallèles. En revanche, cette implémentation est susceptible de ne donner que de faibles résultats lorsque les corpus sont de natures très diverses. Cette méthode n'est donc pas viable sur des corpus à large échelle. Dans notre cas spécifique, voici les résultats que nous obtenons.

Nous proposons une seconde approche, pour lequel nous sommes encore en phase exploratoire, à savoir l'utilisation de modèles massifs multilingues. Pour cette seconde expérimentation, nous prenons le modèle multilingue mBERT, qui connaît le latin, mais pas le grec ancien. Nous affinons le modèle, entre autres, sur des données parallèles que nous avons obtenues grâce aux travaux de Gérard Gréco (Gréco 2008) qui a parallélisé les traductions juxtalinéaires du grec et du latin (essentiellement du corpus homérique et platonicien), mais aussi essentiellement à partir de données que nous avons pu créer, qui ne sont pas encore publiables. Nous accordons au modèle un « learning rate » (capacité d'apprentissage à partir de données d'entraînement) important lors de l'affinage, permettant ainsi au modèle de tirer le meilleur parti des données d'entraînement à l'affinage. Cette méthode, déjà mise en place par d'autres chercheurs pour les langues à faibles ressources (Wang et al. 2020) permet de mettre à profit les équivalences sémantiques déjà acquises du modèle, et de lui fournir de nouvelles connaissances par l'apport de données parallélisées minimales.

Notre première consiste à entraîner des « embeddings » sur les auteurs latins les plus anciens, afin de comparer les résultats d'une même opération sur les auteurs plus tardifs. Nous verrons aussi ce que l'expérience donne sur un entraînement plus global.

Le résultat est sans appel pour MUSE : plus le corpus d'entraînement est faible, plus les résultats sont mitigés. Commençons par les résultats obtenus sur des sèmes *topoi* des corpus philosophiques, à commencer par « ψυχή », « ἀληθής », « ἀρετή », « δόξα » et « λόγος ».

Pour « ψυχή » sur un corpus global nous obtenons comme mots les plus proches « corpus », « anima » et « sensus ». Pour « ἀληθής », toujours sur le corpus global, nous obtenons « verus », « verissime » et « prudens ». Pour « ἀρετή », nous obtenons « sapientia », « beatitas », « virtus ». Pour « δόξα », nous obtenons « veritas », « mediocritas », « falsitas » et « dubitatio ». Enfin, pour « λόγος », nous obtenons « verbum », « sermo » et « disputatio ».

Sur le corpus restreint des premiers auteurs latins de l'époque classique en revanche, les performances sont moins évidentes, mais les correspondances plus spécifiques. Pour « ψυχή » nous obtenons d'abord « animal », puis « mos » et « adsensus ». Pour « ἀληθής », nous obtenons en premier lieu « prudens », puis « verane » et « probus » (« verissime » n'arrive que beaucoup plus tard dans la liste). Le terme « ἀρετή » est déjà moins sujet à variation que les autres termes étudiés, puisque paraissent en premier lieu « virtus », « sapientia » et « divinitas ». Pour « δόξα », nous obtenons « veritas » essentiellement (sous forme d'adverbe ou d'adjectif dans la suite de la liste), et beaucoup plus loin « inprudentia » (mais le caractère potentiellement péjoratif est beaucoup moins présent). Enfin, pour « λόγος », les mots obtenus semblent avoir un caractère plus concret, à savoir « verbum », mais aussi « oratio », « memoratio » et « versutiloquus ».

Sur le corpus restreint des derniers auteurs latins de l'époque classique, « ψυχή » devient « anima », puis « pectura », puis « animus ». Le terme « ἀληθής » est plus marqué encore par ce phénomène de substantification des sèmes philosophiques, puisqu'il obtient très peu de résultats probants tels quels (« obstantia », « iactantia », « negligentia », voire « minutia », alors que le terme « ἀλήθεια » est plus proche des termes « reverentia », « patientia », « cohaerentia » et « verus »). Le terme « ἀρετή » est très proche d'« ἀλήθεια », et a pour plus proches voisins « cohaerentia », « fulgentia », « dignitas » ou encore « claritas ». Le terme « δόξα » prend un tour nettement plus péjoratif, beaucoup plus proche de « iactantia » ou encore « omnipotentia ». Enfin, « λόγος » a désormais comme plus proches voisins « veritas », « verecundia », « modice » et d'autres déclinaisons de « verus ».

Cette méthode tend à montrer que les mots étudiés se distancient peu à peu de leur aspect concret en latin, et tendent peu à peu à l'abstraction, voire à une compréhension beaucoup plus moraliste des termes grecs. Cette première expérience, sur un échantillon relativement restreint, semble adéquate.

Nous pouvons cependant constater que les résultats obtenus via la méthode des « embeddings » contextuels permet une approche plus fine des sèmes obtenus, particulièrement lorsque l'on entraîne les vecteurs sur des empans chronologiques plus précis. Le besoin en termes de données est moins important (sauf à l'affinage du modèle) que celui nécessaire à MUSE, qui obtient systématiquement des résultats moins probants sur des empans chronologiques réduits. Cela tient au fait que les modèles issus de mBERT sont beaucoup plus sensibles aux contextes immédiats, et permettent donc potentiellement une meilleure étude de l'évolution des sèmes dans le temps.

Sur un entraînement sur le corpus global, Le modèle BERT multilingue obtient de très bons résultats, proches de ceux obtenus avec la méthode d'alignement des espaces

vectoriels. Par exemple, pour « ψυχή », nous obtenons « anima » en première position, puis « mens » et « corpus », ou plus loin « spiritus ». Les scores de similarité obtenus sont tous supérieurs à 0.97 (1 étant le score maximal). Pour « ἀληθής », les résultats sont beaucoup plus faibles, voire fautifs sur le corpus global, alors que pour « ἀλήθεια », les résultats redeviennent excellents (avec en premières positions « veritas », « honeste », « recte »). La substantification du terme oriente le sens vers une appréhension morale de la notion. Pour « ἀρετή », le score est encore bon, sans une valeur en dessous de 0.98, avec en premier lieu « veritas », « honestas », « virtus » et « dignitas ». Quant à « δόξα », nous obtenons « voluntas », « opinio », « suspicium » et « vultus ». Enfin, pour « λόγος », les résultats sont légèrement plus faibles mais adéquats, puisque nous obtenons « oratio » en première place, « scribo » et « explico », puis « textus » et « liber » (« verbum » n'arrive qu'en vingtième position).

Mais c'est surtout lors des entraînements par tranche que le modèle BERT semble le plus efficace, avec des termes plus précis (et potentiellement plus adéquats vis-à-vis du corpus, plus limité dans le temps, traité). Par exemple, pour « ψυχή », en première position apparaît « mens », puis « animus », « corpus », « natura » et « homo ». Pour « ἀληθής », les résultats sont davantage probants sur cet empan chronologique, puisque nous obtenons « respondo », « nequio » et « pervenio », mettant davantage l'accent sur la connotation dialogique du contexte dans lequel le terme apparaît, tandis que « ἀλήθεια » conserve « veritas », « honeste », mais aussi « aequus » et « legitimus ». Le vecteur d'« ἀρετή » conserve les plus proches voisins comme « honestas » ou « veritas », mais avec bien davantage de dérivés de « veritas » (comme « vere » ou « verissime »), mais aussi « libertas ». Le vecteur de « δόξα » est aussi relativement précis, mettant en lumière le caractère négatif de la « δόξα », avec des termes comme « suspicium » au premier rang (avec plusieurs dérivés, comme « suspicio »). Pour « λόγος », la connotation de la loi est encore plus présente, avec « oratio », « scribo », « iuris », « argumentum » et « lex ». « Loquor » apparaît en dixième position.

La connotation religieuse de certains des termes étudiés est beaucoup plus marquée lorsque l'on étudie les vecteurs obtenus sur la seconde tranche du corpus latin. Pour « ψυχή », nous obtenons « anima », « persona » et « corpus », mais aussi « spiritus ». Pour « ἀληθής », nous obtenons essentiellement du bruit, dont l'entraînement global a hérité, et les résultats sont plus difficilement interprétables. On note cependant une forte prédominance des termes grecs au sein même du corpus latin pour ce vecteur, avec entre autres « ὀρθῶς », et de termes très connotés négativement, comme « inefficacus » ou « pereo ». Le vecteur d'« ἀλήθεια » est lui aussi nettement marqué par une dimension morale, voire prescriptive, avec une forte présence d'« honestus » (sous plusieurs formes), de « rectus » (lui aussi sous plusieurs formes », de « virtus » et de

« necessarius » (là encore sous plusieurs formes). Le vecteur d'« ἀρετή » correspond quasiment exclusivement à des substantifs à forte connotation morale, comme « virtus », « dignitas », « qualitas », « potestas », « auctoritas », et seuls, dans les dix premiers plus proches vecteurs, se trouvent deux adjectifs, aussi fortement connotés : « bonus » et « aeternus ». Pour « δόξα », les vecteurs proches restent sensiblement les mêmes, à l'exception de « fides », « lux », « imago » et « gratia ». Enfin, pour « λόγος », les scores sont moindres, mais apparaissent néanmoins des termes jusque-là absents des deux autres ensembles de vecteurs, à savoir « liber », « ratio » et, en troisième position, « narratio ».

Il semble donc que les modèles contextuels de type BERT offrent un potentiel important pour l'amélioration de la compréhension de l'évolution des langues. Ils sont capables de mettre en lumière des évolutions fines sur des tranches temporelles relativement restreintes avec peu de données. Notre exemple est réduit, mais l'application à un corpus homogène et ciblé semble cohérente du point de vue des résultats.

Conclusion

Les deux méthodes que nous avons employées pour mesurer les proximités sémantiques multilingues sont encore à l'état de travail en cours. Les deux méthodes employées dans cet article nous semblent à ce jour complémentaires, la première pouvant à ce stade permettre d'enrichir les données pour améliorer la seconde. Mais les perspectives de telles méthodes sont nombreuses, et dépassent largement le cadre de l'étude du latin et du grec. Tout d'abord, le fait que ces méthodes fonctionnent permet de penser qu'il sera plus aisément faisable de créer des corpus parallèles pour l'entraînement de nouveaux modèles. Ensuite, avec l'augmentation de la masse et de la diversité des corpus accessibles, notamment via les progrès de l'HTR (Handwritten Text Recognition), il est raisonnable de penser que nous pourrons obtenir des modèles de plus en plus variés et génériques, propres à être affinés sur des données particulières de manière plus efficace. Enfin, il en découle qu'une meilleure compréhension de l'évolution sémantique entre les langues (avec des études paramétrées sur des tranches temporelles par exemple), y compris les langues rares, sera envisageable grâce à l'entraînement de nouveaux modèles massifs.

Déclaration d'intérêt

Les auteurs déclarent n'avoir aucun conflit d'intérêts relativement à la rédaction et au contenu de cet article.

Contributions

Éditorial

Rédacteurs en chef des numéros spéciaux

Emmanuel Château-Dutier, Université de Montréal, Canada
Barbara Bordalejo, University of Lethbridge, Canada
Roopika Risam, Dartmouth College, United States

Éditeur de section et de mise en page

A K M Iftekhar Khalid, The Journal Incubator, University of Lethbridge, Canada

Rédacteur en chef et éditeur de traduction

Davide Pafumi, The Journal Incubator, University of Lethbridge, Canada

Éditrice de production

Christa Avram, The Journal Incubator, University of Lethbridge, Canada

Références

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu and Xiaoqiang Zheng. 2015. « TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems ». Consulté le 17 décembre 2023. http://download.tensorflow.org/paper/whitepaper2015.pdf.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, et Tomas Mikolov. 2016. « Enriching Word Vectors with Subword Information ». arXiv. Consulté le 17 décembre 2023. https://doi.org/10.48550/arXiv.1607.04606.

Clérice, Thibault, et Jean-Baptiste Camps. 2022. « nlp-pie-taggers: 0.0.40 ». Zenodo. Consulté le 17 décembre 2023. https://doi.org/10.5281/zenodo.3883589.

Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, et Hervé Jégou. 2017. « Word Translation Without Parallel Data ». arXiv. Consulté le 17 décembre 2023. https://doi.org/10.48550/arXiv.1710.04087.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, et Veselin Stoyanov. 2020.

« Unsupervised Cross-lingual Representation Learning at Scale ». arXiv. Consulté le 17 décembre 2023. https://doi.org/10.48550/arXiv.1911.02116.

Crane, Gregory. 1987. « Perseus Digital Library ». Consulté le 17 décembre 2023. http://www.perseus.tufts.edu/hopper/.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, et Kristina Toutanova. 2019. « BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding ». arXiv. Consulté le 17 décembre 2023. https://doi.org/10.48550/arXiv.1810.04805.

Elliott, Tom, Gabriel Bodard, et Hugh Cayless. 2023. « Epidoc: Epigraphic Documents in TEI XML ». Consulté le 17 décembre. https://epidoc.stoa.org/.

Firth, John R. 1957. « A Synopsis of Linguistic Theory, 1930-1955 ». In *Studies in Linguistic Analysis*, dirigé par John R. Firth, 1-31. Oxford: Blackwell.

Gréco, Gérard. 2008. « Éditions Juxtalinéaires. Latin, Grec, Juxta ». Consulté le 17 décembre 2023. http://gerardgreco.free.fr/.

Johnson, Kyle P., Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. 2021. « The Classical Language Toolkit: An NLP Framework for Pre-Modern Languages ». In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, dirigé par Heng Ji, Jong C. Park, et Rui Xia, 20-29. Association for Computational Linguistics. Consulté le 17 décembre 2023. https://doi.org/10.18653/v1/2021.acl-demo.3.

Manjavacas, Enrique, Ákos Kádár, et Mike Kestemont. 2019. « Improving Lemmatization of Non-Standard Languages with Joint Learning ». In *Proceedings of the 2019 Conference of the North*, dirigé par Jill Burstein, Christy Doran, et Thamar Solorio, 1493-1503. Association for Computational Linguistics. Consulté le 17 décembre 2023. https://doi.org/10.18653/v1/N19-1153.

Miaschi, Alessio, et Felice Dell'Orletta. 2020. « Contextual and Non-contextual Word Embeddings: an In-Depth Linguistic Investigation ». In *Proceedings of the 5th Workshop on Representation Learning for NLP*, dirigé par Spandana Gella, Johannes Welbl, Marek Rei, Fabio Petroni, Patrick Lewis, Emma Strubell, Minjoon Seo, et Hannaneh Hajishirzi, 110-119. Consulté le 17 décembre 2023. https://doi.org/10.18653/v1/2020.repl4nlp-1.15.

Mikolov, Tomas, Kai Chen, Greg Corrado, et Jeffrey Dean. 2013. « Efficient Estimation of Word Representations in Vector Space ». arXiv. Consulté le 17 décembre 2023. https://doi.org/10.48550/arXiv.1301.3781.

Pennington, Jeffrey, Richard Socher, et Christopher Manning. 2014. « GloVe: Global Vectors for Word Representation ». In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, dirigé par Alessandro Moschitti, Bo Pang, et Walter Daelemans, 1532-1543. Association for Computational Linguistics. Consulté le 17 décembre 2023. https://doi.org/10.3115/v1/D14-1162.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, et Christopher D. Manning. 2020. « Stanza: A Python Natural Language Processing Toolkit for Many Human Languages ». In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, dirigé par Asli Celikyilmaz and Tsung-Hsien Wen, 101-107. Association for

Computational Linguistics. Consulté le 17 décembre 2023. https://doi.org/10.18653/v1/2020.acl-demos.14.

Reboul, Marianne. 2023. «Code with Pickles for the Embeddings». Consulté le 19 décembre. https://github.com/OdysseusPolymetis/digital_studies.git.

Wang, Zihan, Karthikeyan K, Stephen Mayhew, et Dan Roth. 2020. « Extending Multilingual BERT to Low-Resource Languages ». In *Findings of the Association for Computational Linguistics: EMNLP 2020*, dirigé par Trevor Cohn, Yulan He, et Yang Liu, 2649-2656. Association for Computational Linguistics. Consulté le 17 décembre 2023. https://doi.org/10.18653/v1/2020.findings-emnlp.240.