Connecter les chapitres linguistiques de Programming Historian ?

Matthias Gille Levenson, Celian Ringwald, Marie Flesch, Jennifer Isasi, Sofia Papastamkou, Riva Quiroga, David Valentine

Connecter les chapitres linguistiques de *Programming Historian* ? Premières ébauches d'une table conceptuelle multilingue constituée semi-automatiquement

Matthias Gille Levenson[1], **Célian Ringwald**[2], **Marie Flesch**[3], **Jennifer Isasi**[4], **Sofia Papastamkou**[5], **Riva Quiroga**[6] et **David Valentine**[7]

[1]École nationale des chartes / École Normale Supérieure de Lyon
matthias.gille-levenson@ens.lyon.fr
[2]Université Côte d'Azur, Inria, CNRS, I3S
celian.ringwald@inria.fr
[3]LLF, CNRS–Université Paris Cité
marie.flesch@gmail.com
[4]Pennsylvania State University, USA
j.isasi@psu.edu
[5]Luxembourg Centre for Contemporary and Digital History, Université du Luxembourg
sofia.papastamkou@uni.lu
[6]Pontificia Universidad Católica de Chile
riva.quiroga@uc.cl
[7]Université de Montréal
david.valentine@umontreal.ca

Résumé

Ce travail présente un projet naissant d'alignement automatisé des différents chapitres linguistiques de *Programming Historian*, une revue en ligne didactique proposant des leçons en humanités numériques ainsi qu'un ensemble de traductions de ces leçons, les deux types de publications étant évalués par les pairs. L'objectif principal en est la constitution d'une table conceptuelle multilingue permettant (i) de servir de référence pour les traductions et (ii) de pouvoir relier les différentes versions linguistiques des leçons afin d'être plus efficace sur le plan didactique.

1 Introduction

1.1 Les versions linguistiques de Programming Historian

Programming Historian (aussi PH) est une revue de didactique des humanités numériques qui propose un grand nombre de leçons techniques et didactiques évaluées par les pairs sur des thèmes de sciences humaines qui sont d'une grande diversité (Papastamkou et al., 2021). Si son origine est anglophone, avec un premier chapitre créé en 2012, de nouvelles versions linguistiques sont rapidement proposées (en espagnol en 2016, en français en 2018, en portugais en 2020), afin de faciliter la diffusion de savoirs techniques bien souvent disponibles en langue anglaise, permettant leur utilisation et leur adoption par une communauté plus large que la seule communauté anglophone (Rojas Castro et al., 2019).

Programming Historian est une réussite du point de vue de son audience, qui est aujourd'hui internationale et diverse : sur les trois dernières années le nombre de visiteurs du site est estimé à environ un million par an (Crymble et Im, 2023).

1.2 Les défis de la traduction pour le PH

Dans un paysage des humanités numériques dominé par l'anglais, le multilinguisme et la traduction sont un enjeu majeur de la recherche et de la didactique [1]. La mise à la disposition de ressources dans d'autres langues favorise la diversification de la communauté et l'accessibilité des techniques numériques aux personnes n'ayant pas un bagage linguistique en anglais suffisant ; elle permet aussi, à rebours, une meilleure compréhension de l'anglais technique, omniprésent dans le champ (Galina Russel, 2014) [2].

La question de la traduction des termes techniques des humanités numériques pose toutefois

1. Voir ici Rojas Castro et al. 2019; Fiormonte 2021; Risam 2018. Pour le lien étroit qu'entretiennent didactique et recherche en Humanité Numériques, voir l'étude de Daniel Alves (2021) sur le cas portugais.

2. Voir plus spécifiquement le travail d'Allés-Torrent et Riande (2020) pour une réflexion sur le sujet en prenant comme objet d'étude l'enseignement de la TEI en contexte hispanophone.

des défis importants [3]. Ceux-ci sont dus à l'omniprésence de l'anglais dans les outils et les documentations, à la variation terminologique entre ces outils et aux influences qu'ils exercent sur les traductions, à la variation linguistique liée aux diverses traditions scientifiques qui se rencontrent dans les humanités numériques, ainsi qu'à la variation géolinguistique.

1.3 Objectifs

Le multilinguisme de *Programming Historian* est un élément structurant de la revue depuis 2016, et répond à un besoin de la communauté des lecteurs de la revue [4]. Peut-on se servir de ce multilinguisme pour renforcer la qualité didactique des leçons ? Notre approche [5] est double. Premièrement, elle vise à offrir la possibilité aux personnes parcourant les leçons d'accéder au concept tel qu'il est employé dans la langue de rédaction originale de l'article (souvent l'anglais), et éventuellement dans les autres langues de *Programming Historian* : c'est l'intérêt externe. En second lieu, ce travail aurait un intérêt interne, par le renforcement de l'homogénéité des concepts employés entre leçons, qu'elles soient déjà publiées ou à venir. Les objectifs de notre travail et de notre étude sont donc les suivants :

1. aligner les traductions de concepts techniques et méthodologiques ;
2. documenter les méthodologies de traduction (origine géographique et/ou objectifs pédagogiques du traducteur ou de la traductrice) ;
3. créer un guide de traduction à la fois descriptif (aide à la compréhension pour les lecteur·rice·s de PH) et prescriptif (guide à la traduction, utilisé par les traducteur·ice·s), sous la forme d'une table de concepts alignés dans toutes les langues (pour les leçons effectivement traduites) ;
4. améliorer les traductions, en corrigeant le manque d'homogénéité par exemple.

Notre travail prend donc la forme dans un premier temps d'une table de concordances entre termes techniques, que nous appelons table conceptuelle multilingue. Cette table est une première manière de relier les leçons entre elles. En d'autres termes, il s'agit de passer du modèle de coexistence de leçons dans des versions linguistiques multiples à un modèle de dialogue entre ces mêmes leçons.

La création d'un tel vocabulaire multilingue sera une aide précieuse dans les choix éditoriaux effectués par *Programming Historian* lors de la rédaction d'une leçon ou de sa traduction, mais aussi lors de la phase d'évaluation par les expert·e·s, souvent issu·e·s de différents pays. Ce vocabulaire, ébauché dans cet article permettra de renforcer l'homogénéité des leçons, sera gage de qualité éditoriale, et facilitera le travail des contributeur·trice·s de *Programming Historian*, leur permettant de se concentrer sur un travail de localisation, c'est-à-dire d'adaptation linguistique mais aussi culturelle des leçons au public visé (Isasi et Rojas Castro, 2021; Isasi et al., 2023; Risam, 2018).

Notons ici que notre but ici n'est pas de proposer un vocabulaire qui privilégierait une variété géolinguistique sur une autre et renforcerait sa prééminence (comme le français de France), mais, en restant ouvert à la variation diatopique, de guider les choix de traduction pour des raisons didactiques (Sichani et al., 2019). Il s'inscrit ainsi dans la continuation de la réflexion des autres équipes linguistiques de la revue (Isasi et Rojas Castro, 2021).

Ce travail préparatoire constitue une prototype d'un projet de plus grande ampleur. Il se centre sur la production d'une chaîne de traitement complète, depuis la récupération des leçons jusqu'à la production d'une table multilingue. L'évaluation des

3. On peut penser à la traduction de concepts techniques de programmation (« chaîne de caractères », « boucle », « liste », « dictionnaire », etc.), de termes liés à l'utilisation courante de l'informatique (« fichier », « répertoire », « menu », etc.), mais aussi de termes plus abstraits (« données », « valeur », « ontologie », etc.). Le français opte souvent pour un calque de l'anglais. Certains calques peuvent poser des problèmes à cause de la polysémie du mot source : l'anglais *normalization*, par exemple, qui a plusieurs acceptions à la fois en TAL et en statistique. De plus, l'usage peut hésiter ou varier en fonction des disciplines. C'est notamment le cas de *close reading*, un concept né dans le champ de l'analyse littéraire (Moretti, 2013) que l'on trouve traduit par « lecture proche », « lecture attentive » et « lecture immanente », ou bien simplement utilisé sous forme d'emprunt. La variation géographique ajoute une complexité supplémentaire. Par exemple, l'anglais *cheatsheet* est traduit par *chuleta* en Espagne, *torpedo* au Chili, *machete* en Argentine ou encore *acordeón* au Mexique. Au-delà d'entraîner des problèmes de compréhension, des choix de traduction peuvent paraître très étonnants, voire choquants, pour certains locuteur·trice·s : c'est le cas de la traduction de l'élément TEI *milestone* en espagnol d'Espagne (castillan) par *mojón*, qui signifie « étron » en Amérique du Sud. Dans ces deux cas, l'équipe hispanophone de PH a opté pour des traductions plus consensuelles : *hoja de referencia* pour *cheatsheet*, et *hito* pour *milestone*.

4. Voir le premier ticket mentionnant le besoin d'une version hispanophone de la revue : https://github.com/programminghistorian/jekyll/issues/246.

5. Ce projet est porté par certains membres de l'équipe francophone et hispanophone de *Programming Historian* et a bénéficié de discussions avec des membres des autres chapitres linguistiques de la revue ; il ne porte cependant en aucun cas la voix de *Programming Historian* dans son ensemble.

méthodologies utilisées fera l'objet de productions ultérieures.

2 État de l'art

Il existe depuis plusieurs années un certain nombre de référentiels techniques multilingues qu'il s'agira de faire dialoguer avec la table une fois qu'elle aura été créée et consolidée. Parmi ces référentiels, on trouve notamment :

— la *Taxonomy of Digital Research Activities in the Humanities* (TaDiRAH, Borek et al. 2021), déjà en partie utilisée par PH dans la description des aires thématiques de chaque leçon ;
— `https://glosario.carpentries.org`, un glossaire créé par un projet relativement proche de *Programming Historian*, mais plus fondamentalement axé vers la programmation sans l'approche orientée sciences humaines propre à PH (Pugachev, 2019) ;
— Wikipédia, qui peut aussi être utilisé pour de la traduction technique validée par les communautés de pratique

Il est important de ne pas produire un nouveau standard multilingue de concepts en humanités numériques, mais au contraire de nous aligner sur ce qui existe, sans s'interdire de diverger pour autant en cas de désaccord. Un premier travail de création d'une table de concordance a été effectué par les équipes lusophones et hispanophones de PH, qui propose, à destination interne une fois de plus, une petite table de concordance entre l'espagnol et l'anglais [6].

À côté de ces lexiques, de nouvelles pratiques émergent fondées sur l'apparition récente de grands modèles de langues (LLM, pour *Large Language Models*) extrêmement performants (Min et al., 2023). Ces LLM ouvrent la voie à des solutions auparavant inenvisageables comme l'alignement automatique de grands volumes textuels multilingues, que ce soit au niveau de la phrase ou du syntagme (Artetxe et Schwenk, 2019; Reboul, 2022; Craig et al., 2023; Liu et Zhu, 2023).

6. Une mémoire de traduction technique a aussi été ébauchée par l'initiative « R para la ciencia de datos » : `https://github.com/cienciadedatos/documentacion-traduccion-r4ds/blob/master/orientaciones-traduccion.md`.

3 Chaîne de traitement

3.1 Corpus final retenu

Nous entendons ici par *leçon* l'ensemble des documents composé d'un texte original et de ses traductions. Le corpus de leçons originellement intégré à la chaîne de traitement présentée ci-dessous comptait 255 documents (toutes versions linguistiques confondues). Des problèmes de structuration ont mené à écarter une centaine de documents (ajouts de divisions dans une leçon par exemple) : 152 documents ont pu être transformés en XML-TEI de façon automatisée, ce qui correspond à 75 leçons.

L'étude des traductions étant l'objectif principal de ce travail, les leçons représentées par une version linguistique uniquement ont été écartées. Le corpus final est donc de 37 leçons et de 115 documents. Du point de vue de l'équilibre des langues représentées, toutes les leçons ont leur version originale en anglais (37) ; 32 leçons sont traduites en espagnol, 28 en portugais et 18 en français.

3.2 Prétraitement et structuration

La première étape est la conversion des leçons – rédigées au format MarkDown avec un certaine tolérance pour certaines balises HTML – en TEI, afin de proposer des données mieux structurées, sémantisées et plus aisément manipulables. Ce changement de format permet d'homogénéiser et de standardiser l'information, en plus d'en faciliter la consultation ; il permettra par la suite de relier finement les versions linguistiques une fois l'alignement effectué, à l'aide de pointeurs spécifiques (par ex. `@corresp`). Par ailleurs, la mise à disposition des articles au format XML-TEI (avec documentation *ad hoc*) garantira une plus grande réutilisabilité du corpus de *Programming Historian* pour d'autres projets.

Dans le cadre de notre projet, la structuration en XML-TEI permet de faciliter l'étude et l'alignement des leçons, en sélectionnant les éléments textuels à aligner et ceux à exclure. Dans le second cas de figure, on trouve notamment les blocs de code, moins intéressants, qui peuvent poser des problèmes de tokénisation (la ponctuation étant largement utilisée dans beaucoup de langages différents), et qui contiennent souvent peu de texte traduit.

Un schéma et un ODD (Burnard, 2019) documentent les données produites, avec une restriction du nombre d'éléments autorisés à 47. Cet ODD est

Anglais	Espagnol	Français	Portugais
url , response , and webContent are all variables that we have named ourselves .	url , respuesta y contenidoWeb son variables que nosotros mismos hemos llamado así .	Nous avons nous-même instancié les variables url , reponse et contenu web .	url , response e webContent são todas variáveis nomeadas por nós .

TABLEAU 1 – Extrait d'une des tables d'alignement produites.

publié avec les données [7]. Une fois produit, le corpus est tokénisé au niveau de la phrase en utilisant des marqueurs de ponctuation forts comme le point, le point d'interrogation ou le point d'exclamation. Cette tokénisation est effectuée « à la volée » dans le XML-TEI, ce qui permet par la suite de produire un corpus aligné en XML-TEI.

3.3 Alignement structurel

La structuration peut permettre un premier alignement macroscopique du texte, en partant du présupposé qu'une leçon et ses traductions doivent avoir la même structure. Une traduction très fidèle aura en effet tendance à reproduire la structuration du document source, ce qui permet un alignement préalable division par division, voire paragraphe par paragraphe [8]. Seules les leçons qui correspondent à ce critère d'homogénéité structurelle sont retenus pour la suite de la chaîne de traitement.

3.4 Corpus final retenu

Nous entendons ici par *leçon* le corpus de documents composé d'un texte original et de ses traductions. Le corpus de leçons originellement intégré à la chaîne de traitement comptait 255 documents (toutes versions linguistiques confondues). Des problèmes de structuration ont mené à écarter une centaine de documents (ajouts de divisions dans une leçon par exemple) : 152 documents ont pu être transformés en XML-TEI de façon automatisée, ce qui correspond à 75 leçons.

L'étude des traductions étant l'objectif principal de ce travail, les leçons représentées par une version linguistique uniquement ont été écartées. Le corpus final est donc de 37 leçons et de 115 documents. Du point de vue de l'équilibre des langues représentées, toutes les leçons ont leur version originale en anglais (37) ; 32 leçons sont traduites en espagnol, 28 en portugais et 18 en français.

4 Alignement

4.1 Extraction des concepts

L'outil utilisé pour l'identification des concepts est GliNER (Zaratiana et al., 2024). Ce modèle basé sur BERT a été entraîné de manière à pouvoir extraire des types d'entités définis en amont par l'utilisateur.

GliNER permet la récupération d'entités nommées de tout type à partir de termes définis par l'utilisateur. Nous avons donc choisi les étiquettes ["technical term", "programming", "code", "informatics"] afin de mener l'extraction phrase par phrase, faisant usage de la tokénisation préalable au niveau de la phrase. Plus de 500 entités sont récupérées à partir du corpus de leçons en anglais

Un premier corpus de termes est donc théoriquement directement récupérable à partir des leçons, sans avoir besoin d'alignement à la phrase : il suffit de pouvoir aligner directement les concepts extraits dans les différentes langues. Cependant, produire la table d'alignement permet de compenser l'absence éventuelle d'identification de termes dans une des langues, et c'est la raison pour laquelle cette option a été préférée, outre l'intérêt de pouvoir disposer d'un alignement fin de toutes les leçons.

4.2 Alignement phrase par phrase

Les leçons sont alignées à l'aide de l'outil Bertalign (Liu et Zhu, 2023), qui utilise un grand modèle de langue, par défaut, LaBSE (Feng et al., 2022), et qui ne nécessite pas d'affinage pour produire de bons résultats sur des états de langue contemporains. L'alignement fonctionnant sur des *bitexts*, une phase de traitement [9] doit être appliquée afin de connecter toutes les leçons entre elle avec un pivot,

7. https://github.com/matgille/papier_humanistica_2024/blob/c5815c/data/documentation.

8. Cette première phase a en particulier permis d'identifier des erreurs de structuration courantes dans les leçons, la structuration en format markdown étant propice aux sauts de divisions, car la hiérarchie ne s'exprime qu'à l'aide du nombre de caractères dièse # en début de titre.

9. La méthodologie utilisée consiste en une transformation des paires d'alignement dans un réseau, les noeuds connectés représentant les unités alignées au pivot et donc entre elles.

Anglais	Français	Portugais
the major learning outcome is to compare the time it takes to **manually convert data formats** , compared to doing it with code .	son objectif pédagogique principal consiste à comparer le temps qu'il faut pour **convertir des dates manuellement** avec celui requis lorsque l'on dispose de code pour le faire .	o principal resultado do aprendizado é comparar o tempo necessário para **converter manualmente formatos de dados** , em comparação com fazê-lo com código .

TABLEAU 2 – Une imprécision de traduction dans une des leçons alignées

qui est la leçon originale. Cette phase donne lieu à un premier tableau multilingue où chacune des leçons est alignée au niveau de la phrase : le tableau 1 est tirée de la table d'alignement d'une des leçons, « *Downloading Web Pages with Python* [10] ».

4.3 Alignement des concepts et création de la table

Une fois les phrases alignées une à une et les concepts d'intérêt identifiés, il est nécessaire de passer au niveau intraphrastique pour aligner les concepts identifiés. Le problème principal, comme pour l'alignement de phrases, est l'alignement d'une unité vers plusieurs et de plusieurs vers une (« *one to many, many to one* »), ainsi que le problème des inversions, beaucoup moins important dans l'alignement macrotextuel.

Pour résoudre ce problème, a été utilisé *awesome-align* (Dou et Neubig, 2021), qui permet un alignement non linéaire au niveau du mot, sur chacune des phrases pré-alignées [11] ; le résultat est croisé avec la table de concepts identifiés préalablement, dont le fonctionnement a été décrit plus haut (4.1). De la sorte, la traduction de chaque concept extrait dans la version originale est identifiée dans chacune des traductions de la leçon. De cet alignement au mot, seuls sont extraits et conservés les concepts techniques.

Une autre approche a été testée, à savoir la vérification de la présence des concepts identifiés dans Wikipedia, à des fins de vérification des traductions de termes techniques. Pour récupérer ces « metaliens » nous avons fait appel à DBpedia, qui nous permet d'interroger Wikipédia comme une base de données, en limitant cette exploration à l'anglais et au français. Afin de rendre la recherche plus efficace, a été utilisé DBpedia Lookup [12] afin d'obtenir les pages les plus à même de décrire les concepts extraits par GliNER ainsi que de récupérer les URIs des pages Wikipedia relatifs à ces concepts. Une fois obtenus nous requêtons l'*endpoint* DBpedia [13] qui renvoie les liens `owl:sameAs` interlangues entre les pages Wikipedia précédemment collectées [14].

4.4 Premiers résultats

La phase d'extraction des concept [15] est particulièrement sensible à divers paramètres : le choix de la segmmentation du texte (au niveau du paragraphe ou de la phrase), le choix du paramètre de seuillage de GliNER, mais aussi la version linguistique du modèle (version anglophone ou multilingue du modèle de langue). La version anglophone de GliNER a été privilégiée, étant donné que l'ensemble des leçons retenues ont pour origine une leçon en anglais [16]. Enfin nous avons pu remarquer que notre méthode d'extraction de concept pouvait facilement nous retourner des entités nommées invariables selon la langue (ex : Python, Javascript). Il est évidemment intéressant d'observer que ces unités sont invariables, il serait cependant peut être plus utile de pouvoir à travers notre démarche faire justement ressortir les unités qui sont utilisées de manière divergentes à travers les langues.

La table conceptuelle produite permet par ailleurs d'identifier des erreurs de traduction. Ainsi,

10. L'ensemble des tables d'alignement est à retrouver ici : https://github.com/matgille/papier_humanistica_2024/blob/c5815c/data/alignment_tables.

11. Nous conservons les hyperparamètres par défaut présentés dans l'article.

12. https://github.com/dbpedia/lookup

13. https://dbpedia.org/sparql

14. Un premier cas d'usage basé sur la leçon « Downloading Web Pages with Python » est disponible à l'adresse suivante : https://github.com/matgille/papier_humanistica_2024/blob/c5815c/data/wikipedia_alignment_use_case.html

15. On pourra trouver la table non nettoyée à l'adresse suivante : https://github.com/matgille/papier_humanistica_2024/blob/main/data/aligned_concepts_table.html

16. Le nombre d'entités extraites avec le modèle multilingue est bien moins important, de l'ordre de deux fois moins par rapport à la version monolingue anglaise avec le plus de paramètres.

l'expression « *convert data formats* » est-elle traduite par « convertir des dates » dans la leçon « Introduction to Jupyter notebooks », comme le montre la table 2, ce qui est imprécis, bien que relativement cohérent avec l'objet pris comme exemple dans la leçon. L'identification de cette imprécision a été rendue possible par consultation de la table des concepts alignés, bien que non corrigée et encore très perfectible.

4.5 Conclusions, limites et travaux futurs

La principale limite de l'état du travail réside dans le manque d'une évaluation nous permettant de juger de la pertinence de chacune des étapes de la chaîne de traitement : extraction des concepts, alignement des phrases, alignement multilingue des concepts extraits. La création de corpus d'évaluation semble donc une étape importante qui renforcera la scientificité de nos travaux, et permettra par ailleurs une meilleure sélection des hyperparamètres choisis pour chacune des étapes de traitement du corpus.

Le nettoyage et la discussion sur les entrées à conserver dans la table et les traduction acceptés pour chaque concept devra être le second objectif de ce projet. Outre l'extraction de concepts au cœur de notre projet, le travail effectué dans le cadre de cet article peut facilement servir de base au sein de l'association ou ailleurs à quiconque souhaiterait alignement et la comparer de documents multilingues à différentes échelles. L'alignement des leçons et l'établissement de cette table conceptuelle permet d'envisager plusieurs avancées : premièrement, l'identification des traductions erronées ou imprécises, ou de la traduction hétérogène de concepts (un même concept qui serait différemment traduit entre deux ou plusieurs leçons), ce qui pourra permettre de corriger et d'amender les différentes traductions publiées. En second lieu, il pourra donner matière à une étude amorcée pour le chapitre hispanophone (Isasi et Rojas Castro, 2021) sur les pratiques de traductions, les traducteurs et traductrices – non professionnelles – étant identifiées et leur variété linguistique connue.

Enfin, la table, après validation par les différents chapitres linguistiques de la revue pourra faire office de mémoire de traduction et servir de guide pour la production de nouvelles leçons [17]. Elle pourra dans un second temps être enrichie et transformée en glossaire multilingue proprement dit par l'ajout de définitions dans chacune des langues du *Programming Historian*, ou bien en un thésaurus formalisé via le langue SKOS, permettant d'agréger des concepts hiérarchiquement et de produire des entrées normalisées à partir des concepts déclinés en discours dans les leçons (par une phase de lemmatisation par exemple), ce qui permettrait à ce travail d'être partagé et réutilisé dans d'autres contextes au sein de la communauté en Humanité Numériques.

4.6 Code et données

Toutes les données et scripts sont disponibles en accès libre sur le dépôt github du projet : `https://github.com/matgille/papier_humanistica_2024`.

Bibliographie

Susanna Allés-Torrent et Gimena del Rio Riande. 2020. The Switchover: Teaching and Learning the Text Encoding Initiative in Spanish. *Journal of the Text Encoding Initiative*, 12.

Daniel Alves. 2021. Ensinar humanidades digitais sem as humanidades digitais : um olhar a partir das licenciaturas em história. *EducaOnline*, 15(2).

Mikel Artetxe et Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7 :597–610.

Luise Borek, Canan Hastik, Vera Khramova, Klaus Illmayer, et Jonathan D. Geiger. 2021. Information organization and access in digital humanities: TaDiRAH revised, formalized and FAIR. In *Information between Data and Knowledge*. Werner Hülsbusch.

Lou Burnard. 2019. What is TEI Conformance, and Why Should You Care? *Journal of the Text Encoding Initiative*, 12.

Caroline Craig, Kartik Goyal, Gregory Crane, Farnoosh Shamsian, et David A Smith. 2023. Testing the Limits of Neural Sentence Alignment Models on Classical Greek and Latin Texts and Translations. In *Proceedings of the CHR2023 conference*.

Adam Crymble et Charlotte M. H. Im. 2023. Measuring digital humanities learning requirements in Spanish & English-speaking practitioner communities. *International Journal of Digital Humanities*, 5(2-3) :253–282.

Zi-Yi Dou et Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

17. L'utilisation des scripts présentés dans cet article au sein du processus d'intégration continue de *Programming Historian* pourrait de même être envisagée afin de seconder le travail des traducteur·ices et des personnes chargées de l'évaluation des traductions au cours de la période de production des traduction.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, et Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Domenico Fiormonte. 2021. Taxation against overrepresentation ? the consequences of monolingualism for digital humanities. In *Alternative Historiographies of the Digital Humanities*, pages 333–376. Punctum Books.

Isabel Galina Russel. 2014. Geographical and linguistic diversity in the digital humanities. *Literary and Linguistic Computing*, 29(3) :307–316.

Jennifer Isasi, Riva Quiroga, Nabeel Siddiqui, Joana Vieira Paulino, et Alex Wermer-Colan. 2023. A model for multilingual and multicultural digital scholarship methods publishing : The case of programming historian. In Lorella Viola et Paul Spence, éditeurs, *Multilingual digital humanities*, pages 17–30. Routledge.

Jennifer Isasi et Antonio Rojas Castro. 2021. ¿Sin equivalencia? Una reflexión sobre la traducción al español de recursos educativos abiertos. *Hispania*, 104(4) :613–624.

Lei Liu et Min Zhu. 2023. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2) :621–634.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, et Dan Roth. 2023. Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *ACM Computing Surveys*, 56(2) :30 :1–30 :40.

Franco Moretti. 2013. *Distant Reading*, 1 édition. Verso.

Sofia Papastamkou, Jessica Parr, et Riva Quiroga. 2021. Challenges for Digital Literacy in the Humanities: The Open, Community-Based and Multilinguistic Approach of The Programming Historian. In *NewsEye's International Conference*.

Sarah Pugachev. 2019. What Are « The Carpentries » and What Are They Doing in the Library? *portal : Libraries and the Academy*, 19(2) :209–214.

Marianne Reboul. 2022. *Comparaison semi-automatique des traductions françaises de l'Odyssée d'Homère (1547-1955)*. Classiques Garnier.

Roopika Risam. 2018. *New Digital Worlds. Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Northwestern University Press.

Antonio Rojas Castro, Sofia Papastamkou, et Anna-Maria Sichani. 2019. Three Challenges in Developing Open Multilingual DH Educational Resources The Case of The Programming Historian. In *DH2019 :Complexity*.

Anna-Maria Sichani, James Baker, Maria José Afanador Llach, et Brandon Walsh. 2019. Diversity and inclusion in digital scholarship and pedagogy: The case of The Programming Historian. *Insights*, 32(1).

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, et Thierry Charnois. 2024. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In *NAACL 2024*.