Constitution d'un corpus de référence en orthophonie, issu de la base ISTEX (libres de droits) en langue française et anglaise : rapport technique

Frédérique Brin-Henry (ATILF-CNRS), Clémentine Arnicot (ATILF-CNRS), Sabine Barreaux (INIST-CNRS).

Contact : frederique.henry@atilf.fr

1. Contexte

Ce corpus a été constitué dans le cadre du projet 2019 MOCOLANG-O, subventionné par le pôle scientifique CLCS (Connaissance, Langage, Communication, Sociétés) de l'Université de Lorraine, la Fédération Nationale des Orthophonistes, l'ATILF et le CH de Bar-le-Duc.

Les membres en sont Frédérique Brin-Henry (chef de projet), Rute Costa (NOVA CLUNL, Lisbonne), Sylvie Desprès (LIMICS, Paris), Sabine Barreaux (INIST, Nancy), Anne Dehêtre (FNO, Paris). Clémentine Arnicot (M2 Orthophonie), y a participé en tant que vacataire annotatrice.

Le projet MOCOLANG-O vise la mise au point d'un modèle opérationnel (format OWL) structurant les concepts relatifs aux troubles rencontrés en orthophonie (affectant le langage, la communication et les fonctions oro-myo-faciales). Eminemment interdisciplinaire, il allie orthophonie, linguistique, terminologie et logique de description, et introduit la temporalité comme propriété centrale des concepts. Cette ressource ontologique appelée TemPO (TEMporalité, Pathologie orthophonique, Ontologie) se construit à partir de caractéristiques essentielles issues d'analyses sémantico-syntaxiques précédemment publiées (notamment le projet OrthoCorpus 2015/2017). Le modèle sera éprouvé en le testant avec une terminologie pilote trilingue et un corpus de textes de spécialité issus de la plateforme ISTEX (https://www.istex.fr). En permettant la création et la validation d'une ontologie, il représente une étape d'un projet plus large consacré à la terminologie européenne multilingue en orthophonie, visant à faciliter la communication impliquant les professionnels de santé et leurs patients.

Ce document relate les étapes nécessaires à la constitution du corpus ISTEX /MOCOLANG-O. Ce corpus est destiné à tester la ressource ontologique TemPO, en fournissant des exemples de contextes d'utilisation d'une liste pré-établie de termes diagnostiques dans un corpus élargi et bilingue. Il a été rédigé par F Brin-Henry [*Resp*], et Clémentine Arnicot [*Ann*], et relu par Sabine Barreaux [*IE*].

2. Constitution du corpus français

Une première version d'un corpus en langue française avait été constituée en 2017 dans le cadre du projet OrthoCorpus déjà mentionné. Ce corpus a été mis à disposition en 2018, il comportait alors 39 articles (https://orthophonie-collection.corpus.istex.fr/).

Le corpus augmenté en 2019 dans le cadre de ce projet MOCOLANG-O a été constitué grâce à la requête suivante utilisée pour extraire les articles depuis ISTEX. Ces descripteurs ont été sélectionnés de façon assez large afin de garantir un résultat conséquent.

'(title:(l'orthophoni* d'orthophoni* orthophon* logothérap* logotherap* logopéd* logopèd* logoped* logopaed* "pathologie du langage" "pathologies du langage" "trouble du langage" "troubles du langage") OR abstract:(l'orthophoni* d'orthophoni* orthophon* logothérap* logotherap* "speech therapist" "speech therapists" "speech therapy" "speech therapies" "language therapy" "language therapies" "language

therapist" "language therapists" "speech pathology" "speech pathologies" "speech pathologist" "speech pathologists" "language pathology" "language pathologies" "language pathologist" "language pathologists" logopéd* logopèd* logoped* logopaed* "pathologie du langage" "pathologies du langage" "trouble du langage" "troubles du langage") OR subject.value:(orthophon* logothérap* logotherap* "speech therapist" "speech therapists" "speech therapy" "speech therapies" "language therapy" "language therapies" "language therapist" "language therapists" "speech pathology" "speech pathologies" "speech pathologist" "speech pathologists" "language pathology" "language pathologies" "language pathologist" "language pathologists" logopéd* logopèd* logoped* logopaed* "pathologie du langage" "pathologies du langage" "trouble du langage" "troubles du langage")) AND language:fre'.

Ce corpus comporte 71 articles au début du projet. Trois articles ont été rejetés car ne correspondaient pas à la discipline. Ils étaient relatifs à la logothérapie, technique psychanalytique qui ne relève pas du domaine de l'orthophonie. La liste des descripteurs a dû être ajustée.

Un classeur Excel a été créé, et complété de façon automatique, puis manuelle. L'attribution de métadonnées complémentaires avait pour objectif :

1. De valider les classes thèmes/sous-thèmes utilisées dans la constitution des métadonnées pour le corpus Orthocorpus V2, qui doit paraitre en 2020. La version 1.1 de ce même corpus est rendue disponible sur la plate-forme Ortolang.
2. Les métadonnées ont été obtenues soit par récupération automatique des éléments via l'indexation ISTEX, soit par attribution manuelle.

La section suivante détaille les étapes de production de ces métadonnées.

3. Attribution des métadonnées au corpus français

Les métadonnées ont été extraites soit de façon automatique par [*IE*], soit de façon manuelle par [*Ann*] et validées par [*Resp*]. Le tableau ci-dessous détaille l'ensemble des métadonnées attribuées au corpus :

Tableau 1: métadonnées du corpus

Métadonnée	***Extraction automatique/attribution manuelle***	***Format/modalités***
Identifiant de l'article	Manuelle	ortho_001 et suite
Titre de l'article	Automatique	Titre en français complet
Auteur(s)	Automatique	Les exposants (référence à l'affiliation) sont conservés. Prénoms en entier ou uniquement l'initiale
Affiliation(s)	Automatique	Organisme et adresse
Profession des auteurs	Manuelle	Professions de santé (orthophoniste, médecins) ou psychologue
Auteur orthophoniste ?	Manuelle	Oui /non
Auteurs multiples ?	Manuelle	Oui/non
Pays de l'auteur	Manuelle	Récupéré d'après l'affiliation
Revue ou monographie	Automatique	Titre de la revue ou de la monographie
Editeur	Automatique	EDP Sciences, Lavoisier, Elsevier, Springer (journals)
Type de publication	Automatique	*Journal* ou *book-series*
Type de document	Automatique (Ces catégories ont été proposées par les éditeurs qui fournissent les documents)	*Article, brief communication, other, research article*
Date de publication	Automatique	Année uniquement
Langue(s) du document	Automatique	Français par défaut
Résumé	Automatique	Comprend l'ensemble du résumé français
Public concerné	Manuelle	Professionnel, expert, grand public
Nombre de pages	Manuelle	Repris à partir de l'article récupéré au format pdf

Mots-clés d'auteur	Automatique.	En français puis en anglais. Ex : Hémiplégie ; rééducation ; orthophonie ; langage ; hémiplégie ; adulte ; aphasie ; speech therapy ; language rehabilitation ; hemiplegia ; adult ; aphasia
Thème	Manuelle	Voir section 3.7
Sous-thème	Manuelle	Voir section 3.7
Objet de l'article	Manuelle	données théoriques, matériel et outils, méthodes et techniques de rééducation
Catégories WoS	Automatique	Ex : 1 - social science ; 2 - psychology, experimental ; 1 - science ; 2 - neurosciences ; 2 - behavioral sciences
Catégories Science-Metrix	Automatique	Ex : 1 - health sciences ; 2 - psychology & cognitive sciences ; 3 - experimental psychology
Catégories Scopus	Automatique	Ex : 1 - Life Sciences ; 2 - Neuroscience ; 3 - Behavioral Neuroscience ; 1 - Life Sciences ; 2 - Neuroscience ; 3 - Cognitive Neuroscience ; 1 - Social Sciences ; 2 - Psychology ; 3 - Experimental and Cognitive Psychology
Catégories INIST	Automatique	Ex : 1 - sciences appliquees, technologies et medecines ; 2 - sciences biologiques et medicales ; 3 - sciences medicales ; 4 - neurologie
Score qualité	Automatique	calculé en combinant et en pondérant un certain nombre d'indicateurs de qualité calculés au préalable sur le document (nombre de mots dans le texte intégral, nombre de mots dans l'abstract, version du PDF). Il est noté sur une échelle de 10 points et permet d'avoir une idée rapide de la qualité du texte pour une application en TAL ou TDM.
Version PDF	Automatique	Numéro de version du logiciel Adobe utilisé
XML structuré	Automatique	Oui/non/indéterminé
Identifiant ISTEX	Automatique	Attribué par Istex (ex : 44E1BEC79C6F1053F1B3584C2178FB1C15C2DE39)
ARK	Automatique	Attribué par la plateforme ISTEX afin d'identifier de manière pérenne les documents qu'elle contient.
DOI	Automatique	(Digital Object Identifier) permet l'identification permanente d'un objet électronique publié
PMID	Automatique	Numéro d'identification unique pour chaque article publié référencé dans PubMed

Pour ce qui concerne les items attribués manuellement, les sous-sections suivantes détaillent les décisions prises pour cette attribution.

3.1 Profession des auteurs.

Pour remplir la colonne « profession des auteurs », [*Ann*] a soit récupéré l'information précisée dans l'affiliation, soit elle a procédé à plusieurs recherches Google avec comme mots-clés "nom + prénom/initiale de l'auteur + affiliation" (si cette information était disponible). En cas de doute, [*Ann*] a fait une nouvelle recherche Google pour ces auteurs avec la requête « nom + prénom/initiale de l'auteur + orthophoniste/logopède/logopédiste + nom de la ville ou de l'hôpital ». [*Resp*] a pu répondre à certains doutes grâce à sa connaissance du milieu et son réseau. Suite à cette recherche, si rien n'était trouvé, nous avons considéré que l'auteur n'était pas orthophoniste et avons attribué la valeur « non » dans la colonne « auteur orthophoniste ? ». Le terme « orthophoniste » regroupe à la fois les orthophonistes, les logopèdes et les logopédistes.

Dans la colonne « profession des auteurs », les numéros correspondent aux auteurs selon leur ordre d'apparition dans la colonne « auteurs ».

3.2 Nombre d'auteurs.

[*Ann*] a attribué la valeur « oui » dans la colonne « auteurs multiples » s'il y avait au moins 2 auteurs pour l'article concerné. Cette information est disponible par la récupération automatique de l'information concernant les auteurs de l'article.

3.3 Pays de l'auteur.

L'attribution de la valeur de cette colonne s'est faite grâce à l'affiliation des articles récupérée automatiquement. En cas de doute une requête sur Google a été effectuée. Dans certains cas des auteurs de plusieurs pays ont été impliqués dans la rédaction de l'article, ce phénomène a été retranscrit au moyen d'un + (ex : France+ Belgique).

3.4 Type de document :

Cette métadonnée a été attribuée par les éditeurs. Il semble ici intéressant de préciser cette typologie. La différence entre « research article » et « article » est assez ténue. Un *Research article* (https://content-type.data.istex.fr/ark:/67375/XTP-1JC4F85T-7) rapporte les résultats de travaux de recherche originaux qui sont publiés dans des revues et qui peuvent faire l'objet de communication à des conférence. Un article (https://content-type.data.istex.fr/ark:/67375/XTP-6N5SZHKN-D) fait référence à un article simple publié dans des revues.

3.5 Type de public concerné.

Nous avons fait une différence entre le type de revue (catégorisation automatique) et le public concerné (classement manuel) à la lecture de l'article.

Pour déterminer le type de public concerné, [*Ann*] s'est initialement fondée sur 2 types de public : averti/tout-venant. Puis nous avons décidé de diviser le type de public concerné en 3 catégories : professionnels (généralités qui concernent les professionnels), experts du domaine de l'orthophonie (articles qui s'intéressent à des aspects spécifiquement orthophoniques), et grand public. Après avoir lu les articles et les avoir comparés, nous avons pu formuler des critères permettant de déterminer le type de public concerné.

Tableau 2: critères d'attribution de la classe "type de public concerné"

« Professionnels »	Langage généraliste, approche globale du sujet, langage « médical » mais pas spécifique au jargon orthophonique ou du domaine
« Experts »	Jargon, sujet spécifique du domaine et notamment en orthophonie, compte rendu d'études, méthodologie de recherche en orthophonie
« Grand public »	Pas destiné à un public médical/paramédical, termes explicités

En comparant le type de public concerné et le type de document, on se rend compte que le choix de cette distinction était judicieux, puisqu'il apparaît qu'il n'y a pas de corrélation entre le type de revue et le public concerné : un « research-article » peut aussi bien s'adresser au grand public, à un public d'experts ou à un public de professionnels, c'est avant tout son contenu qui influencera le choix du type de public concerné.

3.6 Nombre de pages.

Pour la colonne « nombre de pages », le nombre indiqué correspond au nombre de pages du titre/résumé aux annexes et bibliographie comprises.

3.7 Thèmes et sous-thèmes.

Cet ensemble de métadonnées a été attribué à partir des propositions déjà élaborées pour la mise au point du corpus OrthoCorpus, actuellement rendu disponible sur la plate-forme Ortolang[1]. Ces classes n'ayant pas été validées de

[1] Analyse et traitement informatique de la langue française - UMR 7118 (ATILF) (2019). *OrthoCorpus* [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, https://hdl.handle.net/11403/orthocorpus/v1.1.

façon experte, leur attribution à des éléments d'un nouveau corpus du même domaine a semblé un bon moyen de valider la pertinence des valeurs possibles pour ces variables.

Du fait des modifications terminologiques récentes, les sous-thèmes « IMOC » (8) et « oro-faciaux » (11) ont été renommés respectivement « paralysie cérébrale » et « fonctions oromyofaciales», dans le but que la terminologie utilisée ici soit en adéquation avec les usages actuels.

Tableau 3: thèmes et sous-thèmes attribuées aux textes du corpus

Thèmes	Sous-thèmes
1. Atteintes congénitales : *relatif à des atteintes et des handicaps congénitaux* 2. Troubles acquis : *relatif à des troubles acquis, intervenant tout au long de la vie* 3. Troubles développementaux : *relatif à des troubles développementaux, se développant au cours de la première partie de la vie* 4. Troubles neurodégénératifs : *relatif à des troubles neurodégénératifs, se développant suite à une atteinte neurologique, généralement dans la deuxième et troisième partie de la vie* 5. Contributeurs : *relatif à des personnalités importantes en orthophonie* *6.* Examens complémentaires : *relatif à des examens complémentaires recommandés par l'orthophoniste ou non* *7.* Recherche : *relatif à l'actualité de la recherche en orthophonie ou dans les domaines liés* 8. N/A : non applicable	1. Aphasies et désordres apparentés 2. Autisme 3. Bégaiement 4. Calcul et raisonnement logico-mathématique 5. Communication 6. Dysarthrie 7. Fonctions cognitives et de haut niveau 8. Paralysie cérébrale 9. Langage écrit 10. Langage oral 11. Fonctions oromyofaciales 12. Surdité 13. Syndromes 14. Troubles associés 15. Déglutition 16. Voix 17. N/A

Nous avons ensuite décidé d'ajouter une classe complémentaire permettant de déterminer **l'objet d'étude de l'article,** qui peut endosser une des 3 valeurs suivantes : données théoriques, matériel et outils, méthodes et techniques de rééducation.

Il a en effet semblé utile d'ajouter cette classe afin de permettre une identification plus complète de chacun des articles en fonction de l'objet de l'article, du thème et du sous-thème du sujet abordé.

Le corpus français augmenté est disponible à l'adresse suivante : https://orthophonie-francaisv2.corpus.istex.fr/ .

4. Constitution du corpus anglais

Parallèlement à la constitution du corpus en langue française et à la validation des métadonnées, [IE] a procédé à la récupération d'articles en langue anglaise.

Les documents en anglais étant beaucoup plus importants dans la plateforme ISTEX, une première requête reprenant simplement les mêmes critères que pour le corpus français a permis de ramener plus de 2000 résultats.

Dans l'objectif de pouvoir aligner deux corpus comparables dans les deux langues, anglais et français, [*Resp*] et [*IE*] ont examiné les différentes possibilités de restreindre les résultats pour le corpus anglais et ont sélectionné 3 sous-corpus proposant des sélections de plus en plus restrictives.

- 1er sous-corpus

Une première sélection a consisté à :

- Supprimer la recherche sur les mots français
- Exclure certains types de document et de publication qui n'étaient pas représentés dans le corpus français (ex : (« book-review »)
- Exclure les documents portant sur la logothérapie ([*Resp*] a procédé à un tri manuel en enlevant les articles relatifs à la logothérapie et ceux clairement médicaux (ex : chirurgie cordectomie).

Ce 1er sous-corpus est composé de 1861 documents et peut être visualisé sur le site suivant : https://orthophonie-anglais1861.corpus.istex.fr/

❖ 2e sous-corpus

Une 2e sélection a consisté à exclure de la recherche les documents auxquels ont été attribuées les catégories Scopus « 1 - physical sciences » et un certain nombre de catégories de niveau 2 qui ne sont pas représentées dans le corpus français. Cette méthode permet de garder les catégories pertinentes ainsi que les documents sans catégories.

Les catégories Scopus ont été choisies parce qu'elles ont la meilleure couverture sur l'ensemble du corpus (85%).

Ce 2e sous-corpus est composé de 692 articles et peut être visualisé sur le site suivant : https://orthophonie-anglais692.corpus.istex.fr/

❖ 3e sous-corpus

Cette 3e sélection a consisté à exclure de la recherche les documents auxquels ont été attribuées les catégories Scopus de niveau 3 qui n'étaient pas représentées dans le corpus français.

Ce 3e sous-corpus est composé de 425 articles et peut être visualisé sur le site suivant : https://orthophonie-anglais425.corpus.istex.fr/

La requête correspondant à ce 3e sous-corpus est la suivante :

Requête : '(title:(logotherap* logotherapeut* logoped* logopaed*) OR abstract:(logotherap* logotherapeut* "speech therapist" "speech therapists" "speech therapy" "speech therapies" "language therapy" "language therapies" "language therapist" "language therapists" "speech pathology" "speech pathologies" "speech pathologist" "speech pathologists" "language pathology" "language pathologies" "language pathologist" "language pathologists" logoped* logopaed*) OR subject.value:(logotherap* logotherapeut* "speech therapist" "speech therapists" "speech therapy" "speech therapies" "language therapy" "language therapies" "language therapist" "language therapists" "speech pathology" "speech pathologies" "speech pathologist" "speech pathologists" "language pathology" "language pathologies" "language pathologist" "language pathologists" logoped* logopaed*)) AND language:eng AND publicationDate:[1969 TO 2016] NOT host.genre.raw:"book" NOT genre:("review-article" abstract "case-report" editorial conference chapter "book-reviews") NOT categories.scopus:("physical sciences" "health professions" "arts and humanities" "nursing" "dentistry" "computer sciences" "agricultural and biological sciences" engineering mathematics "pharmacology, toxicology and pharmaceutics" "business, management and accounting" "immunology and microbiology" "economics, econometrics and finance" ageing anatomy anthropology "applied psychology" Biophysics "Cancer Research" "Cardiology and Cardiovascular Medicine" "Cellular and Molecular Neuroscience" "Clinical Psychology" Communication "Cultural Studies" "Developmental Neuroscience" Embryology "Emergency Medicine" Epidemiology "Family Practice" Gastroenterology "General Neuroscience" "General Psychology" Genetics "Genetics(clinical)" "Geriatrics and Gerontology" "Health Policy" "Health(social science)" Hematology Histology Law "Linguistics and Language" "Neuropsychology and Physiological Psychology" "Obstetrics and Gynaecology" Oncology Otorhinolaryngology "Pathology and Forensic Medicine" "Psychology (miscellaneous)" "Public Health, Environmental and Occupational Health" "Pulmonary and Respiratory Medicine" "Radiology Nuclear Medicine and imaging" "Social Psychology" "Social Sciences (miscellaneous)" "Sociology and Political Science" Surgery) NOT arkIstex:("ark:/67375/WNG-KLQGJ6PX-G" "ark:/67375/6GQ-FL6GVX5C-G" "ark:/67375/WNG-GQX9GF8H-2" "ark:/67375/WNG-14WCN16W-0" "ark:/67375/NDQ-78G6XXV8-5" "ark:/67375/6GQ-GVQLR866-3" "ark:/67375/VQC-6QJDLXP6-C" "ark:/67375/VQC-CM1BHJR9-8" "ark:/67375/VQC-DLQM848T-X" "ark:/67375/VQC-R4BP8RCH-0" "ark:/67375/VQC-R01PR8NT-P" "ark:/67375/VQC-ZV7RNN9X-K" "ark:/67375/VQC-XB60MF2M-C" "ark:/67375/VQC-F0L7T2HW-Z" "ark:/67375/VQC-S6NWQ39B-3" "ark:/67375/VQC-HWSG2N55-8" "ark:/67375/VQC-J0QRW2H9-0" "ark:/67375/1BB-ZS0FGVBV-W" "ark:/67375/WNG-SDGKR545-M" "ark:/67375/WNG-1Q6RH05W-Z" "ark:/67375/WNG-HXG0KBFV-W" "ark:/67375/VQC-BGCM4BXR-Z" "ark:/67375/6GQ-77PB7JCP-D" "ark:/67375/VQC-3L6SH6H6-P" "ark:/67375/VQC-7MVHQTNH-5" "ark:/67375/VQC-729RZ439-0" "ark:/67375/VQC-W0GKJ9JX-3" "ark:/67375/6GQ-S6PG8D77-V" "ark:/67375/VQC-96X0NMZM-4" "ark:/67375/VQC-W900VD0F-W" "ark:/67375/VQC-3WMRRMZ4-J" "ark:/67375/VQC-8TBXJ57G-T" "ark:/67375/VQC-S44G069S-P" "ark:/67375/VQC-91SM49VH-C" "ark:/67375/VQC-V8HTXLPS-S" "ark:/67375/VQC-PZ4F006S-F") '

Plusieurs tentatives de résolution du problème d'alignement entre les articles du corpus en langue française et ceux du corpus en langue anglaise. [*Resp*] et [*IE*] ont été confrontées au fait que les articles n'ont pas les mêmes identités.

4.1 Tentative d'alignement des deux corpus à partir des auteurs uniques

Une première tentative a orienté la réduction du nombre des articles en anglais à partir de la liste des articles rédigés par un seul auteur. Dans le corpus anglais, [IE] a détecté 77 documents issus d'auteurs uniques. Puis [*IE*] a créé une sélection de documents en partant des 77 docs « auteurs uniques », complétés en piochant dans les « auteurs multiples » et en essayant de respecter la répartition souhaitée entre type de publication et le type de documents. Les couleurs correspondent aux ensembles dans lesquels les articles ont été sélectionnés.

4.2 Tentative d'alignement des deux corpus à partir des catégories Scopus

Puis une nouvelle tentative a été faite en tentant d'aligner les catégories Scopus des deux corpus. Pour le corpus en français, la répartition des catégories Scopus montre que les articles appartiennent à 60% dans ces 2 catégories (respectivement 40 et 20%)

1 - Health Sciences ; 2 - Medicine ; 3 - Pediatrics, Perinatology, and Child Health

1 - Health Sciences ; 2 - Medicine ; 3 - Rehabilitation ; 1 - Health Sciences ; 2 - Medicine ; 3 - Orthopedics and Sports Medicine ; 1 - Health Sciences ; 2 - Medicine ; 3 - General Medicine

Pour ce qui concerne les catégories WoS les articles sont catégoriés à 20% en pédiatrie : 1 - science ; 2 - pediatrics

Pour ce qui concerne les catégories INIST les articles appartiennent à 16% en 1 - sciences appliquées, technologies et medecines ; 2 - sciences biologiques et medicales ; 3 - sciences médicales et 17% en SHS (mais 48% sont non étiquetés).

Les catégories Scopus ont semblé les plus pertinentes, cependant leur attribution est tributaire d'un souci d'indexation, donc non expert du domaine, ce qui laisse une marge non négligeable d'erreur.

4.3 Tentative d'alignement manuel

[*Resp*] et [*IE*] ont finalement tenté de procéder à un alignement manuel en partant des 68 textes en français, afin de sélectionner de façon experte une possible correspondance entre un article en anglais et un article en français. Cependant le choix en langue anglaise étant large, la tâche s'avère très difficile sans un parti pris difficile à justifier sur un plan scientifique.

5. Conclusion

En conclusion, il a été décidé de laisser pour le moment les deux corpus en l'état, éventuellement de travailler à l'attribution manuelle des mêmes métadonnées que celles utilisées sur le corpus français afin de permettre un alignement fondé sur le plus grand nombre de métadonnées communes. Il faudra alors envisager de travailler soit sur l'ensemble des travaux retenus (1861), soit les 692 sélectionnés, soit les 425 articles obtenus avec la réduction réalisée sur les articles grâce à l'élimination de certains articles avec des catégories Scopus non présentes dans le corpus français initial. Ce travail pourra être envisagé en 2020 grâce à la demande de financement opérée.

Une autre piste d'alignement consiste à analyser les corpus français et anglais à l'aide d'outils de fouille de textes afin d'obtenir une cartographie des documents qu'ils contiennent et ensuite d'opérer une sélection plus fidèle au profil de chaque corpus. Des outils comme Cillex[2], Gargantext[3] ou CorText[4] sont de bons candidats pour réaliser cette analyse.

[2] https://www.istex.fr/cillex/
[3] https://iscpif.fr/gargantext/
[4] https://www.cortext.net/