Rapport technique de la constitution du corpus OrthoCorpus

Frédérique Brin-Henry, Evelyne Jacquey, Jessika Perignon, Sandrine Ollinger

▶ **To cite this version:**

Frédérique Brin-Henry, Evelyne Jacquey, Jessika Perignon, Sandrine Ollinger. Rapport technique de la constitution du corpus OrthoCorpus. [Rapport de recherche] ATILF. 2019. halshs-02388135v2

HAL Id: halshs-02388135
https://shs.hal.science/halshs-02388135v2

Submitted on 3 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Rapport technique de la constitution du corpus OrthoCorpus

Frédérique Brin-Henry[1,2], Evelyne Jacquey[1], Jessika Perignon[1], Sandrine Ollinger[1]

1 Laboratoire ATILF UMR 7118 université de Lorraine-CNRS, 4, avenue de la Libération, BP 30687 - 54063 Nancy Cedex
2 Centre Hospitalier de Bar-le-Duc, 1 boulevard d'Argonne, CS 50510, 55012 Bar le Duc Cedex

Contact : frederique.henry@atilf.fr

Préambule

Le corpus OrthoCorpus regroupe les articles parus depuis 1997 dans la revue *Rééducation Orthophonique*, revue de référence fondée en 1962 par Suzanne Borel-Maisonny et éditée par Ortho-Edition. Ces articles ont été écrits par des orthophonistes, des linguistes et par d'autres professionnels de santé ou de l'éducation (psychologues, médecins, kinésithérapeutes...), ou par d'autres parties prenantes (représentants d'associations, parents…).

Cette ressource a été constituée dans le cadre du projet du même nom OrthoCorpus (2015-2017), qui vise l'exploitation de corpus de spécialité en orthophonie, dans une perspective d'analyse terminologique. Ce projet a été réalisé avec le concours financier de la Région Lorraine, du laboratoire ATILF (UMR 7118 Université de Lorraine/CNRS), du Centre Hospitalier de Bar-le-Duc et de la Fédération Nationale des Orthophonistes. La ressource est toujours en cours d'augmentation (ajout de tranches d'articles 2017 et 2018 ; ajout de métadonnées). Une première version a été stabilisée pour un dépôt sur Ortolang[1]. Il s'agit de la version V1.1. qui comporte 957 articles parus entre 1997 et 2016, soit 4 millions de tokens.

[1]La plateforme Outils et Ressources pour un Traitement Optimisé de la LANGue (Ortolang) est consultable à l'adresse https://www.ortolang.fr/.

1. Description générale du corpus, version V1.1

Cette version a été déposée et publiée sous Ortolang en décembre 2018. La philosophie générale qui a présidé à la réalisation de ce corpus est de disposer d'un corpus structuré permettant des interrogations selon différents axes, aspects ou facettes. Les différentes facettes selon lesquelles le corpus est interrogeable peuvent se réaliser sous la forme de métadonnées (du corpus et des articles) ou sous la forme d'enrichissements (au niveau de chaque article ou bien au niveau des unités lexicales que celui-ci contient). Au niveau des articles, les facettes interrogeables proviennent de la source des données, par exemple le nom ou le numéro de la revue dans lequel l'article est paru. Les facettes interrogeables proviennent aussi de réalisations manuelles ou automatiques qui ont été mises en œuvre lors de la constitution de cette version du corpus. Il s'agit notamment de métadonnées associées manuellement à chaque article comme le pays d'appartenance de l'auteur principal de l'article, la profession du ou des auteur(s), le type de patient dont traite l'article (enfant ou adulte). Par le biais d'enrichissements automatiques ou semi-automatiques, d'autres facettes sont interrogeables : des éléments de structure au niveau du texte de l'article (introduction, conclusion, corps du texte, mots-clés, titre, etc) et des enrichissements au niveau des unités lexicales de l'article (lemmatisation, étiquetage grammatical).

Les formats dans lesquels est encodé le corpus ont été choisis en fonction des facettes que nous avons voulues structurantes et interrogeables. Le format XML-TEI choisi est un format de stockage et de mise à disposition du corpus. Le format TXM choisi permet l'interrogation de l'ensemble des facettes structurantes du corpus, quelles que soient leur origine (métadonnées ou enrichissements, manuel ou automatique) et leur portée (texte, structures du texte, unités lexicales).

Les données du corpus ainsi que leur description sont accessibles sur l'espace dédié sous Ortolang. Les procédures et les outils qui ont permis de les réaliser peuvent être mis à disposition sur demande à partir d'un espace partagé, interne au projet. Au total, la réalisation de la ressource a mobilisé cinq personnes :

- Frédérique Brin-Henry, responsable scientifique du projet, *RESP* ci-après ;
- Concettina Husson-Giardina, annotatrice, *ANN* ci-après ;
- Évelyne Jacquey, chercheuse dans le projet, *CH* ci-après ;
- Sandrine Ollinger, chercheuse et développeuse, *CH-DEV* ci-après ;
- Jessika Perignon, développeuse, *DEV* ci-après.

La section (2) ci-dessous décrit les grandes étapes qui ont jalonné la constitution du corpus dans sa version V1.1. La section (3) synthétise l'ensemble des étapes ayant

conduit à la version V.1.1 et les classifie selon leur mode de réalisation (automatique, semi-automatique, manuel).

2. Constitution de la version V1.1 : grandes étapes

2.1. Données source

2.1.1. Origine des données

Les données initiales sont les articles reçus par l'éditeur au format Word. Il s'agit de la version « auteur » des textes, choisie pour deux raisons. La première raison est liée au grand nombre d'erreurs de conversion entre d'une part la version de l'article mis en page sur Indesign (le logiciel éditeur de graphisme) vers le pdf, puis une conversion en Word. En effet, des erreurs sur les mots sont possibles en raison de l'OCRisation. De plus l'objet du projet n'était pas de conserver la mise en page éditeur, mais plutôt de nous intéresser au texte de spécialité. La version « auteur » représente donc pour ce projet, malgré un différentiel certain entre celle-ci et la version publiée, une source plus compatible avec nos objectifs. De plus, il nous a paru important de faire la distinction entre les droits sur le contenu (auteurs) et les droits sur la mise en forme (éditeur). Pour la mise en forme, un accord a été négocié entre *RESP* et l'éditeur de la revue. Pour les contenus, chaque auteur a été contacté et son autorisation a été recueillie sauf pour deux auteurs qui ont refusé que leurs articles entrent dans la ressource. Actuellement et depuis 2016, les auteurs reçoivent des instructions dans lesquelles il est mentionné qu'ils cèdent leurs droits à l'éditeur pour cet article.

La revue publie différents types d'articles, dont des éditoriaux et des entretiens – témoignages, qui ont été exclus (sauf à de rares occasions lorsque l'éditorial était substantiel) en raison de leur visée documentaire et du propos considéré comme non scientifique.

Il y a eu deux périodes de récupération des données. Un premier ensemble a concerné les articles parus entre 1997 et 2014. Le second ensemble a concerné les articles parus en 2015 et 2016.

*2.1.2. Homogénéisation des formats Word v**ersion « auteur » d'origine (le « stylage »)***

Les articles initiaux étant sous Word dans une version « auteur », l'ensemble des documents a été uniformisé selon un modèle Word2010 mis au point par *DEV*. À l'issue de cette procédure de stylage, *DEV* assure la transformation des articles vers un format XML de base via OxGarage (section 2.3. et la procédure est décrite en détail sur l'espace interne au projet).

Les fichiers sont tout d'abord renommés selon la convention <annotateur>_<numéro de revue>_auteurs. À l'intérieur du corps de chaque article, les éléments suivants sont mis dans des styles identifiés :

- Nom(s) d'auteur(s) et coordonnées, titre principal en français et en anglais ;
- Mots-clés en français et en anglais ;
- Résumé(s) ;
- Introduction, corps du texte, conclusion, annexes, bibliographie ;
- Remerciements et citations

Les notes de bas de page n'ont pas fait l'objet d'un stylage particulier. Elles ont été détectées et traitées automatiquement via OxGarage.

Les tableaux ont été convertis en image.

La procédure de stylage a pris 4 mois à temps plein. *RESP* et *CH* ont testé la procédure. *ANN* l'a réalisée et achevée dans son intégralité pour les articles parus entre 1997 et 2014. *RESP* et *DEV* ont reproduit cette procédure à l'été 2017 pour les articles parus en 2015 et 2016.

À l'issue de cette étape, l'ensemble des articles sont homogénéisés au format Word2010 selon le modèle du projet. Des procédures de vérification ont permis de corriger les erreurs de doublons ou d'articles manquants dans les classements des articles. Une dernière procédure de vérification a été appliquée afin de comparer les fichiers d'origine avec les fichiers passés au format homogénéisé Word2010, afin de s'assurer de n'avoir oublié aucun fichier.

Les documents du corpus au format homogénéisé Word2010 sont ensuite utilisés de deux manières différentes mais interconnectées :

- la finalisation d'un fichier global de métadonnées (section 2.2) ;
- la mise dans un format XML-TEI des articles en Word2010 selon un modèle défini dans le projet (section 2.3).

Ensuite, l'import TXM du corpus est réalisé à partir du fichier global de métadonnées et des articles au format XML-TEI (section 2.4).

Enfin, le dépôt sur la plateforme Ortolang est préparé en utilisant le fichier global des métadonnées, les articles au format XML-TEI ainsi que le corpus importé sous TXM (section 2.5).

2.2. Métadonnées des articles

Les métadonnées des articles ont été rassemblées dans un unique classeur Excel qui a été vérifié et complété manuellement par *RESP* à l'issue d'une procédure d'extraction semi-automatique mise au point par *DEV* (voir section 2.4). Il est actuellement maintenu par *RESP* et *DEV*. La dernière version à jour de ce fichier est conservée dans l'espace interne au projet.

Ce fichier de métadonnées sous Excel est exploité pour produire les métadonnées des articles au format XML-TEI (voir section 2.3.3.2) et celles des articles au format TXM (voir section 2.4). Le tableau ci-dessous synthétise les exploitations du fichier Excel de métadonnées pour les deux formats dans lesquels le corpus sera mis à disposition sur la plateforme Ortolang.

Champ du fichier Excel	Format XML-TEI (header TEI)	Format TXM, « propriété de structure », structure « text »	Signification du champ
« id »	Nom du fichier xml	champ « id » avec suppression de l'extension « .xml »	Identifiant du fichier correspondant à l'article traité
« Auteur Orthophoniste »	Valeur de l'élément teiHeader | fildeDesc | sourceDesc | note [@type = 'auteurOrthophoniste']	propriété « auteurorthophoniste »	Indication du fait qu'au moins un des auteurs de l'article est orthophoniste ou qu'aucun ne l'est
« Auteur Multiple »	Valeur de l'élément teiHeader | fildeDesc | sourceDesc | note [@type = 'auteurMultiple']	propriété « auteurmultiple »	Indication du fait que l'article ait été écrit par un auteur unique ou par plusieurs auteurs
« pays »	Valeur de l'élément teiHeader | fildeDesc | sourceDesc | note [@type = 'paysAuteur']	propriété « pays »	Indication du pays d'appartenance institutionnelle de l'auteur principal tel qu'il est mentionné dans l'article

« numero »	**Valeur de l'élément teiHeader | fildeDesc | sourceDesc | series | biblScope |@unit = 'numero']**	**propriété « numéro de revue »**	**Numéro de la revue dans lequel l'article a été publié**
« annee »	**Valeur de l'élément teiHeader | fildeDesc | sourceDesc | biblStruct | monogr | imprint | date**	**propriété « annee »**	**Année de parution de l'article**
« titre du numero »	**Valeur de l'élément teiHeader | fildeDesc | sourceDesc | biblStruct | monogr | title [@level = 'm']**	**propriété « titredunumero »**	**Indique le titre du numéro dans lequel l'article a été publié**
« rubrique »	**Valeur de l'élément teiHeader | fildeDesc | sourceDesc | note [@type = 'rubrique']**	**propriété « rubrique »**	**Indique de quelle rubrique éditoriale relève l'article parmi les 8 existant dans la revue**
« enfant ou adulte »	**Valeur de l'élément teiHeader | fildeDesc | sourceDesc | note [@type = 'enfantOuAdulte']**	**propriété « enfantouadulte »**	**Indique si l'article traite d'enfants ou d'adultes**
« titre de l'article »	**Valeur de l'élément teiHeader | fildeDesc | sourceDesc | biblStruct | analytic | title**	**propriété « titredelarticle »**	**Titre de l'article**
« auteur »	**Valeur de l'élément teiHeader | fildeDesc | sourceDesc | biblStruct | analytic | author**	**propriété « auteur »**	**Auteur(s) de l'article**

En plus des métadonnées ci-dessus, *RESP* a procédé à une indexation des articles en catégories, thèmes et sous-thèmes. Cette indexation a été validée en 2019 grâce à un travail effectué dans le cadre du projet MOCOLANG-O[2], et l'ensemble de ces métadonnées pourront être intégrées dans la prochaine version du corpus.

2.3. Passage des articles du format Word2010 au format XML-TEI conforme au modèle du projet

Ce passage est réalisé en trois grandes étapes. La première consiste à passer les articles du format Word2010 à un format XML bien formé (section 2.3.1). La seconde

[2] Projet 2019 financé par le CLCS-Université de Lorraine, la Fédération Nationale des Orthophonistes, ATILF et le CH de Bar-le-Duc

consiste à initialiser le fichier global des métadonnées en extrayant ce qui peut l'être à partir des articles en XML ou déduit en fonction des numéros de revue sur le site de celle-ci (section 2.3.2). Cette première extraction de métadonnées est ensuite enrichie manuellement afin d'aboutir au fichier global de métadonnées. La dernière étape s'appuie sur les deux premières pour passer les articles au format XML-TEI conforme au modèle du projet (section 2.3.3).

2.3.1. Transformation des articles vers un format XML bien formé via OxGarage

Cette étape de traitement fait passer les articles du corpus du format Word 2010 conforme au modèle du projet à un format XML bien formé en s'appuyant sur les styles Word associés dans l'étape précédente. Pour cela, *DEV* utilise la plateforme en ligne OxGarage[3] qui permet de passer un fichier du format docx à un format XML bien formé. Deux précisions sur cette conversion :

- elle ne peut se faire que fichier par fichier ;
- si l'article docx contient une image, la conversion produit une archive zip. Dans ce cas, il est nécessaire d'ouvrir l'archive, de récupérer le document XML et de le renommer selon la convention décrite dans la procédure globale.

À l'issue de la conversion de l'ensemble des fichiers, ceux-ci sont validés avec l'environnement Oxygen selon le modèle d'OxGarage. La validation peut signaler des erreurs dans la balisage de sortie. Beaucoup d'erreurs sont liées :

- à la mise en forme « cachée » de Word, impossible à détecter facilement dans la source ;
- au stylage non conforme : oubli de stylage de certaines parties (nom d'auteur, adresse…) ;
- à des tableaux non transformés en image.

Pour corriger les erreurs liées aux ajouts de mise en forme de Word, il faut modifier le schéma sortieOxGarage et revalider. Pour les erreurs liées au stylage, il faut intervenir dans le Word et repartir à l'étape de conversion via OxGarage jusqu'à validation de tous les fichiers.

3 La plateforme est un service WEB accessible à l'adresse https://oxgarage2.tei-c.org/. Ce service a été développé par Sebastian Rahtz qui le met à disposition de la communauté sous licence GPL3.0. L'ensemble de la procédure est open-source et accessible sous GitHub : https://github.com/sebastianrahtz/oxgarage

À l'issue de cette étape de traitement, les fichiers sont dans un format XML bien formé et conforme aux recommandations de la TEI telles qu'elles sont mises en œuvre dans le modèle de sortie XML de la plateforme de conversion.

2.3.2. Première extraction des métadonnées et préparation du fichier Excel des métadonnées

Une procédure semi-automatique permet d'extraire certaines métadonnées accessibles dans les articles et de récupérer ainsi une première version. Cette première version est vérifiée par *RESP* et *DEV*.

Cependant, les métadonnées accessibles dans les articles ne sont pas complètes. En particulier, il y manque les informations propres à chaque numéro de la revue : l'année, le mois, le titre du numéro, le numéro et le titre de l'article. Ces informations peuvent être trouvées sur le site de la revue. Afin de stocker ces informations venant du site de la revue dans les métadonnées de chaque article XML, pour chaque numéro, une règle de transformation spécifique est codée dans une feuille de transformation XSL qui va s'enrichir au fur et à mesure de l'avancement du traitement des numéros de revue contenant les articles du corpus. À l'issue de cette récupération, la feuille de transformation ainsi produite peut être appliquée sur l'ensemble du corpus ou sur une partie de celui-ci seulement, par exemple à des fins de vérification.

La transformation XSL produit un fichier CSV (tableau texte) qui est daté et envoyé à *RESP*. Ce fichier constitue une version intermédiaire du fichier global de métadonnées. *RESP* le complète ensuite manuellement pour ajouter d'autres informations qui ne sont ni accessibles dans les articles ni sur le site de la revue :

- la présence ou non de plusieurs auteurs ;
- le domaine de spécialité du ou des auteur(s) : orthophoniste ou non ;
- la catégorie, le thème et le sous-thème de l'article[4] ;
- la rubrique du journal à laquelle appartient l'article ;
- si l'article traite d'enfants ou d'adultes.

2.3.3. Mise au format XML-TEI pivot du projet OrthoCorpus

Cette étape est entièrement réalisée sous Oxygen. Elle se subdivise en deux tâches :

- nettoyage de la structure XML des articles ;

[4] Dans la version V1.1 actuellement déposée sur la plateforme Ortolang, les métadonnées « thème » et « sous-thème » sont provisoirement masquées car en attente de validation par Resp qui en a une expertise fine.

- extraction des métadonnées en conformité avec les recommandations de la TEI et passage du format XML bien formé issu d'OxGarage au format XML-TEI du projet OrthoCorpus.

2.3.3.1. Nettoyage et réécriture

L'uniformisation des formats Word des auteurs a été réalisée au sein de l'environnement Word2010. Comme beaucoup d'environnements compilés de ce type, des informations parasites sont insérées et on ne les voit apparaître que lorsqu'on travaille ensuite sur les fichiers produits dans un format explicite (par exemple dans l'environnement Oxygen). Ces informations parasites sont supprimées.

Certains attributs nécessitent un nettoyage voire une réécriture, notamment ceux des éléments .

Les tâches de nettoyage et de réécriture sont réalisées automatiquement sous Oxygen via l'utilisation d'une feuille de transformation XSL qui a été élaborée progressivement par *DEV* au fur et à mesure des styles/attributs/éléments à supprimer, nettoyer ou réécrire.

Dans ce format « propre », l'ensemble de l'information (bibliographies, annexes, styles de mise en forme) est conservé afin de pouvoir archiver une version complète des articles. Dans la version mise à disposition sur la plateforme Ortolang, bibliographie, annexes, notes de bas de page, tableaux et figures sont supprimés.

2.3.3.2. Extraction des métadonnées de chaque article et validation avec le modèle du projet OrthoCorpus

À partir de la version XML « propre » des fichiers du corpus, la mention des auteurs des articles est supprimée dans les noms des fichiers correspondant aux articles et les fichiers sont renommés selon la convention suivante :

- chaque nom de fichier est préfixé par « oc » (pour OrthoCorpus) suivi d'un nombre incrémenté de 1 pour chaque nouvel article (à partir de 0001).

Cette phase de renommage est réalisée tout d'abord au sein du fichier global des métadonnées puis pour chaque fichier du corpus.

À partir du renommage des fichiers, les métadonnées enrichies au format Excel sont injectées dans chaque fichier XML « propre ». Le fichier des métadonnées est d'abord transformé en un fichier XML via la plateforme OxGarage à nouveau. Puis le fichier des métadonnées correspondant en XML est utilisé par une feuille de transformation XSL pour associer à chaque fichier du corpus les métadonnées qui lui correspondent.

Enfin, l'ensemble des fichiers est validé selon le modèle XML-TEI défini au sein du projet.

2.4. Import sous TXM

CH-DEV a rédigé une documentation qui se trouve sur l'espace commun destiné aux membres du projet.

L'import du corpus sous TXM[5] utilise le module XML/w + CSV. Ce module importe des fichiers XML et exploite un fichier de métadonnées appelé metadata.csv.

Le fichier metadata.csv est obtenu par transformation du fichier metadata2018.xsl. Conformément aux recommandations de TXM, il est encodé en UTF-8, chaque champ est encadré de guillemets doubles et séparé des autres par une virgule.

Lors de l'import, la segmentation lexicale est réalisée avec les paramètres par défaut. Une annotation en partie du discours et en lemmes est réalisée à l'aide de Treetagger[6] et du modèle pour le français distribué par l'Université de Munich. Une feuille de style est appliquée. Elle permet deux opérations :

- suppression des teiHeader de chaque fichier et de toutes les notes (balises et contenu) ;
- suppression de certaines balises (mais pas de leur contenu), pour alléger la structure de chaque article.

Pour personnaliser l'édition HTML des articles du corpus, le nombre de mots par page est augmenté à 100 000 et un fichier CSS est édité. La taille du titre de chaque article est ainsi rendue plus grande que celle des titres de sections et les couleurs de la présentation des métadonnées en tête d'article est modifiée.

À l'issue de l'import, les métadonnées du fichier metadata.csv apparaissent comme des propriétés de la structure « text » et les attributs des éléments du XML apparaissent comme des propriétés de la structure « div ». Ces propriétés permettent une interrogation complexe. Elles peuvent être utilisées pour la constitution de sous-corpus (par exemple le sous-corpus des introductions d'articles), le partitionnement du

---

[5] TXM est une plateforme modulaire de textométrie open-source, développé par Matthieu Decorde, Serge Heiden, Sébastien Jacquot, Alexei Lavrentiev et Bénédicte Pincemin et téléchargeable à l'adresse http://textometrie.ens-lyon.fr.

[6] TreeTagger est un étiqueteur morphosyntaxique développé par Helmut Schmid à l'Institute for Computational Linguistics de l'Université de Stuttgart. Il est téléchargeable à cette adresse : https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

corpus (articles traitant des adultes versus des enfants) et des interrogations avancées (recherche de toutes les occurrences de « bilan » dans des introductions d'articles traitant d'enfants). Pour leur part, les annotations apparaissent comme des propriétés des unités lexicales.

Pour interroger le corpus importer sous TXM, la documentation mémo CQL fournit de nombreuses indications sur la manière de bâtir les requêtes.

2.5. Dépôt sur Ortolang

Un dépôt de ressource sur la plateforme Ortolang est un moment important dans la constitution d'une ressource. En plus de la ressource elle-même qui aura été préalablement corrigée et vérifiée, il est nécessaire de définir les modes d'accès à la ressource déposée et de rédiger, en anglais et en français, une présentation de cette ressource. Ensuite, la plateforme Ortolang propose un formulaire, en anglais et en français, qui permet de renseigner les métadonnées pour la ressource déposée. Enfin, la plateforme Ortolang propose une pré-visualisation d'échantillons de la ressource. Il est nécessaire de veiller à ce que cette pré-visualisation soit conforme aux objectifs et attendus du déposant.

La version V1.1 qui est directement accessible depuis la page d'accueil du projet sur la plateforme Ortolang a été déposée en décembre 2018. Elle fait suite à un dépôt antérieur en juillet 2018. Sur l'espace dédié au projet dans la plateforme Ortolang, les deux versions co-existent mais une seule des deux est visible par défaut depuis la page d'accueil. Nous avons choisi la version V1.1 car celle-ci est le résultat de corrections par rapport à la version V1.

2.5.1. Vérification et corrections de la ressource

Dans la version V1, il existait deux versions XML du corpus : l'une pour le dépôt sur Ortolang, l'autre pour l'import dans TXM. En vue du dépôt de la version V1.1, les deux versions XML du corpus ont été fusionnées en une seule par *DEV* afin de n'avoir plus qu'une seule version mise à jour et un seul format pivot. À l'occasion de cette fusion, deux types de corrections ont été mis en œuvre :

- identification et suppression des fichiers en double ;
- correction d'erreurs résiduelles dans la version XML-TEI du corpus correspondant à des erreurs natives (par exemple les notes de bas de page incluses dans les titres) ou liées à la procédure de stylage.

Les métadonnées des articles ont fait l'objet de deux tâches de correction et d'homogénéisation (*RESP*, *DEV*, *CH-DEV* et *CH*) :

- Partant du constat que plusieurs versions du fichier global de métadonnées existaient, deux versions seulement ont été conservées. Une version séparée, à jour et complète a été stockée pour archivage dans l'espace commun destinés aux membres du projet. Une seconde version, amputée des métadonnées jugées non suffisamment abouties (les indications de « thème » et « sous-thème »), a été également déposée. Dans cette version, les noms d'auteurs se trouvent dans une colonne spécifique qui est utilisée notamment lors de l'import sous TXM pour affichage avec la fonctionnalité « Édition » ;
- Du fait des consignes de la procédure de stylage, en particulier la consigne demandant de respecter la casse dans l'indication des noms d'auteurs pour bâtir le nom de fichier d'un document, les noms de fichier dans la ressource telle que déposée dans sa version V1 présentaient de nombreuses irrégularités et un certain nombre d'erreurs. Pour homogénéiser les noms de fichier au sein de la ressource d'une part, et pour gérer la pré-visualisation des échantillons telle qu'elle est fournie sur la plateforme Ortolang d'autre part, le nommage des fichiers par la concaténation des noms d'auteurs a été abandonné. Il a été remplacé par une procédure qui associe à chaque fichier un nom préfixé par « oc » suivi d'un nombre « _i » incrémenté de 1 pour chaque nouveau document de la ressource. Cette modification du nommage des fichiers a été reportée dans le fichier global des métadonnées.
- Enfin, les titres des articles comportant des erreurs de casse ont été manuellement corrigés par *RESP*.

Dernière étape : une nouvelle version du corpus au format TXM a été générée par *CH-DEV*. Cette nouvelle version du corpus au format TXM exploite le corpus XML-TEI corrigé ainsi que le fichier global de métadonnées.

2.5.2. Droits d'accès à la ressource

Les données sont mises à disposition sous licence CC-BY-NC-ND 3.0[7] pour les membres de l'enseignement supérieur et de la recherche (ESR). Cette mise à disposition a été étendue aux orthophonistes chercheurs ou d'autres acteurs du domaine de l'orthophonie qui sont intéressés par la ressource mais qui ne sont pas nécessairement référencés dans l'annuaire de l'ESR. Les membres de ce groupe sont gérés et inscrits nominativement par *RESP* en accord avec l'éditeur de la revue.

Les droits d'accès sont mentionnés dans les métadonnées de la ressource ainsi que dans les métadonnées de chaque document.

[7] http://creativecommons.org/licenses/by-nc-nd/3.0/deed.fr

2.5.3. Gestion de la pré-visualisation sur la plateforme Ortolang

Un squelette TEI-P5 du teiHeader des articles a été créé et mise en place par *DEV* spécifiquement pour le dépôt sur Ortolang de la version V1.1 de la ressource. Ce squelette permet de faire en sorte que le titre apparent de l'article soit le nom du projet accompagné des personnes référentes selon leur(s) rôles au lieu que ce soit le titre authentique (celui fournit par l'auteur) qui apparaisse.

Ce choix vient de notre volonté de protéger la propriété de l'éditeur. L'espace OrthoCorpus est un espace de stockage de données et ne doit en aucun cas être un site de consultation alternatif à celui de la revue qui a fourni les articles. Grâce au squelette défini spécifiquement pour le dépôt, le titre authentique de l'article apparaît dans une police qui ne permet pas de l'identifier comme tel via son apparence. Ce titre authentique n'est alors détectable que par son contenu.

Afin d'alléger la pré-visualisation, le squelette assure que les différentes métadonnées concernant chaque article (titre, auteur, numéro de revue, titre du numéro, année, éditeur de la revue, droits de consultation et de téléchargement, diffuseur de la ressource) sont dans le teiHeader mais n'apparaissent pas dans l'aperçu rapide.

2.5.4. Présentation de la ressource et métadonnées sur la plateforme Ortolang

Au cours du premier semestre 2018, *RESP* a rédigé une présentation en français et en anglais qui constitue la page d'accueil de la ressource sous Ortolang. Au cours de cette même période, les métadonnées de la ressource ont aussi été renseignées.

Ces éléments de présentation ont été revus pour le dépôt de la version V1.1 en décembre 2018.

3. Synthèse et classification des étapes réalisées pour constituer la version V.1.1

Le tableau ci-dessous reprend les étapes décrites pour la parution de la ressource (version V1.1), en donnant une indication temporelle pour chacune d'entre elle, et s'il s'agit d'un procédé automatique ou manuel.

Tâche	Type de réalisation	Durée (mesurée ou estimée)	Commentaire, remarque, précision
Récupération des articles au format Word version « auteur » de la revue *Rééducation Orthophonique*	**automatique puis classement manuel**	**Délai de rigueur de deux ans avant la récupération des articles parus**	**Présence de doublons, articles mal classés dans les dossiers (transmis par années)**

Création d'un modèle Word2010	**manuel**		**Réalisation progressive et testée au fur et à mesure**
Application du modèle Word2010 aux articles de la période 1997 - 2014, 841 documents (1ère phase), puis aux articles de la période 2015 - 2016, 113 documents (2ème phase)	**manuel**	**7 mois ETP (cumulés)**	**Environ 15 à 20 min par article, 10 articles en moyenne par revue, 4 revues par an.**
Transformation des articles au format Word2010 en XML via OxGarage	**semi-automatique**		**Opération assez chronophage dans la mesure où la conversion se fait fichier par fichier. D'autre part, la création et la mise à jour du modèle de validation est progressive et a évolué au fur et à mesure des erreurs rencontrées**
Extraction, homogénéisation et vérification des métadonnées (phase 1)	**manuel et semi-automatique**	**3 mois**	**Vérification et mise à jour indispensable (notamment pour les classes thèmes et sous-thèmes)**
Création d'un squelette XML des métadonnées	**manuel**		**il y a eu plusieurs versions successives jusqu'à celle qui régit le format des articles actuellement déposés**
Association des métadonnées aux articles (phase 2)	**automatique avec corrections manuelles**		
Import dans TXM et paramétrage différencié de l'édition HTML sous TXM	**automatique et manuel**		**A permis de détecter quelques erreurs résiduelles dans les métadonnées**
Préparation du dépôt sous Ortolang	**manuel**		
1. Rédaction des documents de présentation	**manuel**	**2 mois**	
2. Constitution et vérification des métadonnées	**manuel**	**3 mois**	**L'ajout de nouveaux textes oblige à un réexamen complet des métadonnées**
3. Vérification de la visualisation rapide des articles via la plateforme Ortolang	**manuel**		
4. Vérification de l'utilisabilité et de la conformité de la version TXM du corpus	**manuel**		