Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non-uniforme ?

Alix Chagué, Victoria Le Fourner, Manuela Martini, Eric Villemonte de la Clergerie

Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non-uniforme ?

Alix Chagué[1], Victoria Le Fourner[12], Manuela Martini[3], Eric Villemonte de la Clergerie[1]

[1] ALMAnaCH, Inria, Paris
[2] Laboratoire ICT, Université Paris Diderot, Paris
[3] LARHRA, Université Lumière-Lyon 2, Lyon

Introduction

Le processus d'industrialisation a induit des mutations inédites dans les conditions des travailleurs et travailleuses, avec des effets durables sur l'organisation du travail et les niveaux de vie des ménages des couches ouvrières impliquées dans des activités productives en pleine mutation. Parmi les débats classiques sur les causes et les conséquences de la première industrialisation, l'historiographie récente a insisté sur l'importance de l'apport des femmes et des enfants aux revenus de leurs ménages (Humphries 2010).

Après un long silence historiographique, le rôle joué par les femmes dans le développement industriel est désormais largement reconnu (Humphries et Sarasúa 2012 ; van Nederveen Meerkerk 2006 ; van Nederveen Meerkerk et Schmidt 2008). Dans l'un des secteurs clefs de la première industrialisation, l'industrie textile, elles sont présentes dans toutes les phases du processus productif. Pourtant, les données sont lacunaires lorsqu"il s'agit de s'interroger sur leurs rémunérations, les emplois du temps, leur travail domestique et de les comparer à ceux des hommes qui travaillent dans le même secteur. Cette absence de données concerne en particulier la France, à l'instar de l'Europe occidentale et méridionale, à l'inverse du Nord de l'Europe (Ågren 2018 ; Ogilvie 2013). Notre programme de recherche a pour but d'alimenter un débat international foisonnant sur ces questions, auquel nous souhaitons apporter des matériaux documentaires et des résultats inédits.

Présenté en 2016 à l'Agence Nationale pour la Recherche (ANR), sous le titre « *Rémunérations et usages du temps des femmes et des hommes dans l'industrie textile en France, de la fin du XVIIe siècle au début du XXe siècle* », abrégé en « TIME US », notre programme de recherche a débuté en 2017 et se terminera dans le courant de l'année 2020. Sous la coordination de Manuela Martini, il rassemble les membres de plusieurs laboratoires de recherche (outre le LARHRA[1], ALMAnaCH[2],

1 « Laboratoire de Recherche Historique Rhône-Alpes » (UMR 5190) ; Universités Lumière-Lyon 2, Jean Moulin-Lyon 3, Grenoble-Alpes, ENS de Lyon, CNRS ; Lyon. Disponible sur http://larhra.ish-lyon.cnrs.fr/
2 « Automatic Language Modelling and Analysis & Computational Humanities » ; Inria ; Paris.

ICT[3], TELEMMe[4], IRHiS[5] et le Centre Maurice Halbwachs[6]), formant une équipe composée d'historiens et historiennes, de sociologues, de spécialistes du traitement informatique des documents historiques et de spécialistes du traitement automatique des langues.

Le travail mené suppose de rassembler une quantité importante de données et de les organiser pour les rendre exploitables dans le cadre des recherches conduites en histoire économique et sociale, en histoire de la famille et du genre, en histoire des conflits du travail et de la culture des classes populaires, et plus largement par les sciences sociales concernées par des approches longitudinales sur ces questions.

Le programme a donc pour objectif de rendre accessibles des données sérielles issues principalement de sources qualitatives pour évaluer le travail rémunéré et non rémunéré, domestique et extra-domestique des femmes dans l'économie du textile et son industrie. Pour la période concernée, l'attention est spécifiquement portée sur les activités recoupant tout le processus de fabrication de produits textiles : de la filature à la confection, en incluant l'assemblage des pièces. Nous ne nous intéressons pas à la distribution des produits.

Deux approches sont adoptées de manières complémentaires :

- la première, micro-qualitative, implique une analyse empirique très approfondie des contextes historiques de production des sources utilisées ;
- la seconde, quantitative, s'ancre davantage dans le champs des humanités numériques en mettant en relation historiens et historiennes et spécialistes des outils et méthodes informatiques.

C'est cette seconde approche dont nous présentons la méthodologie établie pour la constitution du corpus de textes numériques ainsi que les résultats obtenus. Elle repose sur la collaboration étroite de l'équipe ALMAnaCH avec un nombre plus restreint de partenaires du projet issus du LARHRA ou d'ICT. Dans cette approche, nous souhaitons extraire le texte contenu dans des fichiers images afin de produire des données en langage naturel sous la forme de fichiers XML, structurés suivant le standard de la TEI. Ces fichiers doivent rendre compte de la structure logique du contenu ainsi que d'éléments d'annotation sémantique. Par la suite, une interface de requêtage permettra aux chercheurs et chercheuses du projet d'interroger le corpus, tout en le mettant à disposition pour d'autres usages potentiels. Enfin, dans le cadre du travail mené par l'équipe ALMAnaCH, nous explorons la possibilité d'automatiser les tâches de traitement et en interrogeons la pertinence, c'est-à-dire la faisabilité, le gain ou la perte de temps qu'ils induisent ainsi que la qualité des données en sortie.

Nous présentons, sous la forme d'un retour d'expérience, la chaîne de traitements mise en place pour la composition du corpus de textes. Elle mêle des tâches automatisées, semi-automatisées et manuelles en suivant cinq étapes : collecte des doubles numériques, segmentation, transcription, uniformisation et enfin annotation.

3 « Laboratoire Identité, Culture, Territoire » ; Université Paris Diderot ; Paris.
4 « Temps, Espaces, Langages, Europe Méridionale - Méditerranée » (UMR 7303) ; Aix-Marseille-Université, CNRS, MMSH ; Aix-en-Provence.
5 « Institut de Recherches Historiques du Septentrion » (UMR 8529) ; Université de Lille, CNRS ; Lille.
6 CNRS, ENS, EHESS ; Paris.

Collecte des doubles numériques

Le corpus rassemblé est un mélange d'archives économiques et d'archives juridiques, manuscrites et imprimées et produites à différentes époques. Dans un premier temps, nous avons souhaité utiliser des sources déjà numérisées résultant de campagnes de numérisations menées par d'autres institutions. Numelyo[7], la bibliothèque numérique de la ville de Lyon, permet le téléchargement des numérisations des microfilms de la presse lyonnaise du XIXe siècle. Nous avons collecté un total de 390 fichiers PDF par son biais. L'Université de Toronto a publié sur Internet Archive[8] les numérisations de treize volumes appartenant à la série des « Ouvriers des Deux Mondes ». Près de 8 000 fichiers couleurs au format JP2 sont ainsi disponibles. En dépit de la très bonne qualité de ces fichiers, l'ensemble contient des erreurs de numérisation ou des images jugées inutiles pour le projet, qu'il a donc fallu trier.

Fig. 1: Quelques erreurs de numérisation

En parallèle, lors des dépouillements menés dans les centres d'archives des régions de Lyon et Paris, des doubles numériques inédits, essentiels pour le projet, ont été collectés. Initialement, cette numérisation visait à permettre la transcription manuscrite des sources. Lors des expérimentations menées pour outiller l'équipe en vue de l'automatisation de la transcription, il est apparu nécessaire d'établir une véritable méthodologie pour la prise de vue afin d'obtenir des images de meilleure qualité. Il s'agissait notamment d'améliorer le cadrage, l'éclairage et la définition des images ; ces deux premiers paramètres étant cruciaux pour la réussite de la transcription automatique. En conséquence, nous nous sommes dotés d'un dispositif de prise de vue portatif, la *scanTent* (Projet READ et al. 2018), et d'une plate-forme de stockage partagé afin de faciliter et de garantir la mise en commun des données, en l'occurrence le service *Sharedocs*[9]. De plus, nous avons établi des règles pour la prise de vue, l'organisation des dossiers d'images et la constitution des métadonnées issues de la phase de numérisation. Pour l'ensemble du projet TIME US, nous avons ainsi rassemblé près de 10 000 photographies.

7 Disponible sur https://numelyo.bm-lyon.fr/.
8 Disponible sur https://archive.org/index.php.
9 Disponible sur https://sharedocs.huma-num.fr/.

Ces photographies n'ont pas toutes été traitées dans le cadre de la deuxième approche sur laquelle nous nous concentrons. En effet, nous avons distingué une sélection de cinq sous-ensembles couvrant toute la période du projet sur laquelle ont porté nos expérimentations numériques :

1. Les **contraventions à la police des arts et métiers de la ville de Lyon** prennent la forme de registres de contraventions, produits entre 1667 et 1781[10]. Ils contiennent un texte manuscrit peu structuré. Quoique dépouillés sur la base d'un carottage décennal pour le projet global, nous nous sommes concentrés sur les registres de 1760.
2. La **presse ouvrière lyonnaise** permet de reconstituer une partie des activités du Conseil des Prud'hommes de Lyon dans la mesure où les archives du Conseil ont disparu. Nous avons identifié neuf titres de journaux[11] dans lesquels sont parus, entre 1831 et 1851, en discontinu, des comptes rendus de séances visant en premier lieu à informer et à garder une trace de la jurisprudence. En fonction de la qualité de l'impression, le texte est parfois très bruité, rarement structuré de manière homogène d'un titre à l'autre et d'un numéro à l'autre.
3. Les **minutes du Conseil des Prud'hommes de Paris pour les tissus** (Lemercier 2007) correspondent aux comptes rendus de séances rédigés par le secrétaire de la section du Conseil des Prud'hommes créé en 1847 pour délibérer sur les contentieux dans le textile. Nous nous sommes intéressés aux années 1847-49, 1858, 1868 et 1878[12]. Ce sont des textes manuscrits très structurés, contenant beaucoup d'informations et dont le scribe change peu.
4. Les **monographies familiales de Le Play** (Cardoni 2012 ; *Les monographies de famille de l'École de Le Play (numéro spécial)* 2000) sont des enquêtes publiées par la Société Internationale des Études Pratiques d'Économie Sociale entre 1851 et 1908, sous les titres *Les Ouvriers des deux mondes* et *Les Ouvriers européens*. Le texte, imprimé, est très régulier dans sa mise en forme, mais il contient de nombreux tableaux à la mise en page difficile à appréhender informatiquement.
5. Enfin, les **rapports de police de la préfecture de Lyon** sont un ensemble de rapports rédigés à la suite d'observations sur les mouvements ouvriers à Lyon, en particulier à l'occasion des grèves qui touchent les industries de la soie à la fin de l'année 1894[13]. Ce sont des documents sans structure homogène.

10 Archives municipales de Lyon, HH 214 à 267.
11 *L'Avenir* (1846-1847) ; *L'Écho de la Fabrique* (1831-1834) ; *L'Écho de la Fabrique de 1841* (1841-1845) ; *L'Écho des ouvriers* (1840-1841) ; *L'Écho des travailleurs* (1833-1834) ; *L'Écho de l'industrie* (1845-1846) ; *L'Indicateur* (1834-1835) ; *Tribune prolétaire* (1845-1850) ; *La Tribune lyonnaise* (1834-1835).
12 Archives départementales de Paris, D1 U10 379, 386, 396 et 405, respectivement.
13 Archives départementales du Rhône, 9 M5.

Ensemble	Nature de l'écriture	Mode de collecte	Nombre d'images	Méthode de transcription	Nombre d'images transcrites	Achèvement
Contraventions à la police des arts et métiers de la ville de Lyon	manuscrit	photo.	2 525	manuelle	1 093	43 %
Presse ouvrière lyonnaise	imprimé	téléch.	520	semi-automatique	520	100 %
Minutes du Conseil des Prud'hommes de Paris	manuscrit	photo.	3 439	semi-automatique	1 254	35 %
Monographies familiales de Le Play	imprimé	téléch.	6 500	automatique	6 500	100 %
Rapports de police de la préfecture de Lyon	manuscrit	photo.	451	semi-automatique	126	27 %

Fig. 2: Vue d'ensemble sur le traitement des sous-corpus

Extraction du texte

À partir des numérisations, nous produisons un texte en langage naturel auquel sont appliqués des outils d'analyse syntaxique et textométrique permettant une lecture distante du corpus (Moretti 2013). La mise en place d'une série d'étapes de traitement est nécessaire pour cela. Chacune de ces étapes nous a conduit à établir des approches heuristiques différentes, adaptées aux documents et aux contraintes rencontrées. Par exemple, exception dans notre méthodologie, pour des raisons humaines, l'ensemble des contraventions a été transcrit manuellement, à l'aide d'un traitement de texte. Cette méthode est coûteuse et non reproductible mais produit en général une transcription sans faute. Conformément à nos objectifs, les quatre autres ensembles ont été traités en suivant trois phases de traitement : 1) segmentation des images, 2) transcription, 3) uniformisation du texte extrait.

La segmentation désigne la reconnaissance des zones de texte sur une image, leur typage et leur ordonnancement. Elle découle de l'analyse de la mise en page (*layout analysis*). Nous avons utilisé les solutions disponibles dans l'interface du logiciel Transkribus[14], qui propose deux solutions :

1. La première est une implémentation du logiciel de transcription automatique FineReader[15]. Il présente de bons résultats pour la détection des zones, le typage et l'ordonnancement. Dans Transkribus, FineReader n'est entraîné que pour le texte imprimé et il n'est pas possible de dissocier segmentation et transcription.
2. La seconde est développée par le *Computational Intelligence Technology Lab* (CITLab) de

14 Disponible sur https://read.transkribus.eu/transkribus/.
15 Disponible sur https://www.abbyy.com/fr-fr/finereader/.

l'Université de Rostock (Strauß et al. 2018). Elle propose plusieurs modèles de segmentation adaptés à des mises en pages particulières, comme les cartes postales ou la presse. Même si l'option CITLab fonctionne bien pour les manuscrits, le typage et l'ordonnancement qui en résulte n'est pas suffisant pour les documents complexes.

Le résultat de cette segmentation nécessite parfois des corrections manuelles pour supprimer les faux positifs et compenser les faux négatifs en ajoutant des zones ou en re-définissant les coordonnées d'une zone. Il a ainsi fallu longuement corriger le résultat de la segmentation de la presse ouvrière, que la mise en page en multi-colonne rend difficile, d'autant plus qu'elle est aggravée par la qualité moindre de certaines numérisations. Même dans ce cas extrême, la correction manuelle de la segmentation automatique est plus rapide et moins fastidieuse qu'une segmentation entièrement manuelle.

Les stratégies adoptées pour la transcription ont varié en fonction de la nature des sources et de la qualité des images collectées. Pour les deux ensembles manuscrits, plusieurs modèles d'HTR (*Handwritten Text Recognition*, ou reconnaissance de texte manuscrit) ont été entraînés, atteignant des taux d'erreurs situés entre 19 % et 5,2 %, selon les données fournies. Deux modèles sont particulièrement satisfaisants :

1. Le premier (Prud'hommes_1858_M4+) est entraîné sur les minutes du conseil de Prud'hommes de Paris pour l'année 1858. C'est le meilleur modèle obtenu.
2. Le deuxième (Comb_French_Admin_XIX_M3+) résulte d'une tentative de généraliser l'entraînement en combinant des données issues des rapports de police et des années 1847-49 et 1858 des prud'hommes parisiens.

Pour la presse, FineReader permettait de produire rapidement la transcription des 520 pages à moindre coût en utilisant un modèle pré-entraîné. Enfin, pour les monographies, nous avons utilisé Kraken[16], un logiciel en ligne de commande permettant d'entraîner notre propre modèle d'OCR. Avec peu de données, nous sommes parvenus à un modèle (Model_OD2M) très efficace. En couplant la segmentation issue de Transkribus et la transcription produite avec Kraken, nous avons traduit la totalité des 6 500 pages, tout en contrôlant le jeu de caractères utilisé, ce qui facilite la phase suivante d'uniformisation du texte.

Logiciel	**Modèle**	**Taux d'erreur (CER)**	**Quantité de vérité terrain**
HTR - Transkribus	Prud'hommes_1858_M4+	5,2 %	4 577 lignes
HTR - Transkribus	Comb_French_Admin_XIX_M3+	8,8 %	20 025 lignes
OCR - Kraken	Model_OD2M	2,2 %	1 300 lignes

Fig. 3: Présentation des modèles de transcription

16 Disponible sur http://kraken.re/.

Après deux ans d'exploration méthodologique, de dépouillement, de collecte de numérisations et d'extraction de texte, près de 9 500 images sur les 13 450 concernées ont été transcrites automatiquement, soit près de 70 % du corpus. Quelle que soit la méthode de transcription choisie, des fichiers XML TEI sont produits à ce stade, soit par l'intermédiaire de l'API de Transkribus (Transkribus s. d.), soit par le biais de nos propres scripts de transformation de texte brut ou d'interaction avec Kraken[17].

Uniformisation des données textuelles

Uniformiser les fichiers XML TEI et les données textuelles qu'ils contiennent permet d'aligner nos ressources afin de compenser les écarts découlant des outils et méthodes d'extraction employés tout en conservant les spécificités formelles de chaque source. L'enjeu de cette uniformisation est d'obtenir un corpus homogène compatible avec les outils d'analyse du langage déployés par la suite. Le texte doit être corrigé, les graphies alignées et la structure de chaque ensemble modélisée puis implémentée dans le schéma TEI.

La régularisation des graphies touche en premier lieu les abréviations, qui sont développées, et les dates. Afin de reconstituer le flux des phrases et des paragraphes, les marques de césures, qui sont transcrites, doivent être effacées et les césures résolues. La stratégie adoptée pour cela repose sur un système de règles et d'expression régulières. Pour les césures, le meilleur scénario de résolution est déterminé en fonction des autres formes rencontrées dans le reste des documents. Un premier passage résout une grande partie des cas, les autres seront pris en charge dans le cadre du module de correction automatique en cours de développement à l'automne 2019.

La structuration consiste à détecter et formaliser la hiérarchie de l'information contenue dans un texte. Elle peut être identifiée à partir d'un certain nombre de *features*, parmi lesquelles des indices typographiques, des informations de mise en page (indentation, position sur la page) ou encore des formules récurrentes. Cette tâche commence par une analyse des textes en vue de modéliser la structure des informations. Pour les ensembles peu structurés, nous restons au niveau des unités documentaires (par exemple, les rapports ou les affaires), composés parfois d'un titre ou d'une en-tête, de paragraphes et/ou d'une ou plusieurs signatures. L'ajout des balises se fait par l'intermédiaire de scripts basés sur des systèmes de règles. Dans l'ensemble des monographies, nous repérons en outre les éléments liés à l'édition papier, comme les en-têtes, la pagination et les notes de bas de pages, dont la mise à l'écart permet de recomposer les paragraphes interrompus par un changement de page. La structuration facilite la navigation au sein des ensembles et permet leur éditorialisation. En outre, elle rend possible de concentrer les efforts d'annotation sémantique sur les portions que nous savons susceptibles de contenir les informations ciblées par le projet.

Annotation sémantique

L'annotation sémantique vise à repérer dans le texte les éléments d'information que nous souhaitons extraire et comparer dans l'optique de la recherche historique menée. L'élaboration de cette couche sémantique suppose une collaboration étroite entre les spécialistes de disciplines historiques et

17 « ExportFromTranskribus », « LSE-OD2M », « TEITransformation », disponibles sur https://gitlab.inria.fr/almanach/time-us.

informatiques. Même si la structure de chaque source diffère, les informations que nous souhaitons en extraire sont similaires et doivent pouvoir être repérées grâce à un encodage commun. L'enjeu est donc de parvenir à établir un modèle d'annotation qui tient compte de la diversité des formes que prennent les informations. Cette modélisation est intégrée dans le schéma de la TEI, sous la forme d'une ODD chaînée (Burnard 2016), qui permet de moduler la spécificité des structures de chaque ensemble et l'unité de l'annotation sémantique.

Nous souhaitons repérer trois catégories d'informations, prenant la forme de segments de taille très variables :

- les informations liées aux personnes ;
- les entités et segments liés au travail et aux rémunérations ;
- les entités et segments liés à l'expression du temps.

Pour établir un schéma TEI compatible avec la totalité du corpus, nous partons d'un ensemble puis nous procédons par élargissements progressifs. A l'automne 2019, notre schéma d'annotation sémantique fonctionne correctement avec les ensembles de la presse ouvrière, des comptes rendus de séance du Conseil de Prud'hommes parisien et les rapports de police. Il n'a pas encore été confronté aux deux ensembles restants. Durant cette phase d'élaboration, des portions de texte sont annotées à la main afin de valider les choix de modélisation. Une fois le modèle stabilisé, ces annotations serviront à fournir des données d'entraînement pour automatiser la tâche à l'aide de l'analyseur syntaxique FrMG (Clergerie et al. 2009 ; Morardo et Clergerie 2014), développé à Inria. Pour l'exploration du texte, d'autres outils comme SEM (Dupont 2017) ou TXM (Heiden 2010) sont envisagés en complément.

A ce stade, une première phase de traitements linguistiques a été conduite sur une partie seulement du corpus (les prud'hommes de Lyon et de Paris ainsi que les monographies), représentant un total de 180 098 phrases, à l'aide de la chaîne de traitement du français développée par l'équipe ALMAnaCH et en particulier à l'aide du *parser* FRMG. Les premiers résultats sont encourageants. Les taux de couverture (par phrase entière recevant une analyse complète) vont de 68 %, pour le texte extrait des contraventions, à 91 %, pour la presse ouvrière, et de 78 % à 88 % pour les monographies. L'utilisation d'outils de recherche et de traitements des erreurs devrait conduire à une amélioration de ces taux.

La prochaine étape serait d'utiliser les résultats de l'analyse syntaxique pour acquérir des connaissances dans le domaine concerné par le projet TIME US. Il s'agit en particulier d'extraire des termes et des expressions multi-mots, comme « chef d'atelier » ou « paire de bas », mais aussi de construire des réseaux sémantiques basés sur les hypothèses distributionnelles d'Harris (les mots sémantiquement proches ont tendance à apparaître dans des contextes semblables, ici syntaxiques). Cette étape montre déjà des concepts intéressants pour le domaine :

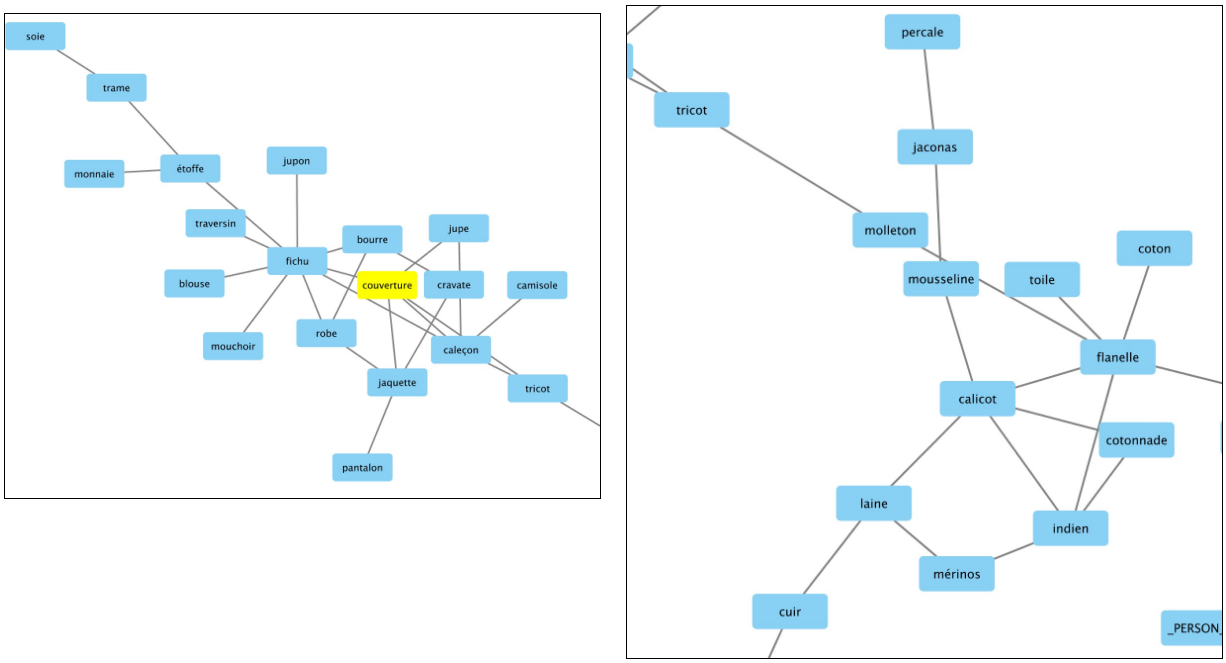

Fig. 4: Réseaux sémantiques généré à partir de l'analyse du corpus

Il est ensuite prévu que soient utilisés certains de ces concepts du domaine de connaissance pour trouver des chemins syntaxiques entre eux. Ces chemins permettront ainsi, par itération, de détecter de nouveaux concepts et de nouveaux chemins, mais aussi de trouver des occurrences où les sous-ensembles du corpus sont liés. Par exemple, les premières expériences révèlent un lien entre « sieur », « somme » et « franc » :

```
(<root> (<subject> (<agent> sieur/nc)) payer/v (<object>
(<patient> somme/nc) (<N2> de/prep (<N2> (<patient>
franc/nc)))))
```

Ce chemin est présent 12 fois dans le corpus, par exemple, dans la phase « Le sieur Tocannier payera la somme de 40 fr ». Notons que sont aussi observées et capturées des variantes de cette relation de paiement.

Une version web du logiciel TXM a été implémentée[18] pour permettre aux historiens et historiennes qui ne sont pas familiers avec les techniques de TAL d'avoir un premier aperçu du corpus. Le portail web permet de consulter le corpus en mode lecture et de réaliser des explorations textométriques, en synergie avec les technologies de corpus actuelles (notamment CQP). Une interface web mieux adaptée à l'expérience utilisateur sera développée de manière à présenter les résultats de l'analyse syntaxique et les réseaux sémantiques. Cette nouvelle interface reprendra les fonctionnalités de TXM et permettra à l'utilisateur de télécharger des sous-parties du corpus en XML TEI ou en texte brut[19].

18 Disponible sur : http://cref.paris.inria.fr/txm/

19 Pour un aperçu de cette future interface, il est possible de visiter la page « Exploring a parsed French Wikipedia », qui permet d'explorer un *dump* du Wikipedia français, disponible sur : http://alpage.inria.fr/frwiki/

Bilan

Chaque ensemble constituant le corpus du projet TIME US a permis d'explorer un large champ de possibilités techniques. In fine, nous réussissons bien à constituer un corpus qu'il est formellement possible de traiter avec un seul et même outil d'annotation sémantique. Il faut toutefois avoir conscience que l'automatisation de ces traitements suppose d'adopter une méthodologie qui repose sur une différente répartition du temps entre les étapes du processus de traitement des sources. Elle suppose aussi de la part des chercheurs et chercheuses une formation aux outils numériques. Nul doute que la répétition de ce genre de protocole permettra d'éviter les écueils chronophages, réduisant le coût technique et humain de tels projets. Finalement, à la question de savoir si l'automatisation permet un gain de temps, nous répondons affirmativement, sans oublier que nous avons obtenu un corpus textuel pérenne, distribuable, et que la communauté pourra exploiter et connecter à d'autres ressources dans le futur.

Bibliographie

Ågren, Maria. 2018. « Making Her Turn Around: The Verb-Oriented Method, the Two-Supporter Model, and the Focus on Practice ». *Early Modern Women: An Interdisciplinary Journal* 13 (1) : 144-152.

Benjamin Kiessling. 2015. « kraken ». http://kraken.re/.

Burnard, Lou. 2016. « ODD Chaining for Beginners ». http://teic.github.io/PDF/howtoChain.pdf.

Cardoni, Fabien. 2012. « Aux sources du budget domestique selon Le Play » Les Études Sociales : 11-46.

Chagué, Alix. 2018. « Constituer un corpus pour la fouille de texte - de la transcription des documents d'archives à l'annotation : exploration d'une méthodologie par l'ANR Time Us ». Mémoire de master. Technologies numériques appliquées à l'histoire. Paris : École nationale des chartes.

Clavert, Frédéric. 2014. *Vers de nouveaux modes de lecture des sources*. FYP EDITIONS. http://orbilu.uni.lu/handle/10993/34980.

Clergerie, Éric Villemonte de la. 2005. « From Metagrammars to Factorized TAG/TIG Parsers ». Dans *Proceedings of the Ninth International Workshop on Parsing Technology*, 190–191. Parsing '05. Stroudsburg, PA, USA : Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1654494.1654516.

Clergerie, Éric Villemonte de la, Benoît Sagot, Lionel Nicolas et Marie-Laure Guénot. 2009. « FRMG: évolutions d'un analyseur syntaxique TAG du français ». https://hal.inria.fr/inria-00553260.

Dupont, Yoann. 2017. « Exploration de traits pour la reconnaissance d'entités nommées du Français par apprentissage automatique ». Dans *19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017)*, édité par Hélène Flamein et Yannick Parmentier. Orléans. http://taln2017.cnrs.fr/wp-content/uploads/2017/06/actes_RECITAL_2017_2.pdf#page=52.

Heiden, Serge. 2010. « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement ». https://halshs.archives-ouvertes.fr/halshs-00549764.

Humphries, Jane. 2010. « The First Industrial Nation and the First "Modern" Family ». Dans *Gender Inequalities, Households and the Production of Well-Being in Modern Europe*, édité par Tindara Addabbo, 41-58. Ashgate Pub Co.

Humphries, Jane et Carmen Sarasúa. 2012. « Off the Record: Reconstructing Women's Labor Force Participation in the European Past ». *Feminist Economics* 18 (4) : 39-67. https://doi.org/10.1080/13545701.2012.746465.

Le Fourner, Victoria. 2019. « Étude de la structuration automatique et de l'éditorialisation d'un corpus hétérogène, l'exemple des sources du conseil des prud'hommes pour le textile du XIXe siècle ». Mémoire de master. Technologies numériques appliquées à l'histoire. Paris : École nationale des chartes.

Lemercier, Claire. 2007. « Juges du commerce et conseillers prud'hommes face à l'ordre judiciaire (1800-1880). La constitution de frontières judiciaires ». Dans *La justice au risque des profanes*, édité par Hélène Michel et Laurent Willemez, 11-27. Paris : PUF.

Les monographies de famille de l'École de Le Play (numéro spécial). 2000. Vol. 131–132. Les Études sociales.

Morardo, Mikaël et Éric Villemonte de la Clergerie. 2014. « Towards an Environment for the Production and the Validation of Lexical Semantic Resources ». Dans *The 9th edition of the Language Resources and Evaluation Conference (LREC)*. https://hal.inria.fr/hal-01005464/document.

Moretti, Franco. 2013. *Distant Reading*. Londres, New York : Verso.

Nakamura-Delloye, Yayoi et Éric Villemonte de La Clergerie. 2010. « Exploitation de résultats d'analyse syntaxique pour extraction semi-supervisée des chemins de relations ». Dans *17e Conférence sur le Traitement Automatique des Langues Naturelles - TALN 2010*. Montréal, Canada. https://hal.archives-ouvertes.fr/hal-00511541.

Nederveen Meerkerk, Elise van. 2006. « Segmentation in the Pre-Industrial Labour Market: Women's Work in the Dutch Textile Industry, 1581–1810 ». *International Review of Social*

History 51 (2) : 189-216. https://doi.org/10.1017/S0020859006002422.

Nederveen Meerkerk, Elise van et Ariadne Schmidt. 2008. « Between Wage Labor and Vocation: Child Labor in Dutch Urban Industry, 1600-1800 ». *Journal of Social History* 41 (3) : 717-736.

Ogilvie, Sheilagh. 2013. *A Bitter Living: Women, Markets, and Social Capital in Early Modern Germany*. Oxford : Oxford University Press.

Projet READ, Computer Vision Lab, Université technique de Vienne et Université d'Innsbruck. 2018. « The ScanTent ». https://scantent.cvl.tuwien.ac.at/en/.

READ. s. d. « Transkribus ». Consulté le 6 novembre 2019. https://read.transkribus.eu/transkribus/.

Rude, Fernand, Ernest Larousse et Édouard Dolléans. 1969. *L'insurrection lyonnaise de novembre 1831: le mouvement ouvrier à Lyon de 1827-1832*. Paris : Editions Anthropos.

Sagot, Benoît et Éric Villemonte de La Clergerie. 2006. « Error Mining in Parsing Results ». Dans *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 329–336. Sydney, Australia : Association for Computational Linguistics. https://doi.org/10.3115/1220175.1220217.

Sewell, William Hamilton. 1980. *Work and revolution in France: the language of labor from the Old Regime to 1848*. Cambridge ; New York : Cambridge University Press.

Stedman Jones, Gareth. 1983. *Languages of class: studies in English working class history, 1832-1982*. Cambridge ; New York : Cambridge University Press.

Strauß, Tobias, Max Weidemann, Johannes Michael, Gundram Leifert, Tobias Grüning et Roger Labahn. 2018. « System Description of CITlab's Recognition & Retrieval Engine for ICDAR2017 Competition on Information Extraction in Historical Handwritten Records ». *arXiv:1804.09943 [cs]*. http://arxiv.org/abs/1804.09943.

« Time Us ». s. d. GitLab. Consulté le 19 novembre 2019. https://gitlab.inria.fr/almanach/time-us.