Articuler archives manuscrites, connaissances marionnettiques et approche linguistique

L'exemple du fonds Chesnais

Stéphane Riou

La chaire ICiMa - Innovation Cirque et Marionnette est un dispositif de recherche porté conjointement par le Centre national des arts du cirque (Cnac) et l'Institut International de la Marionnette (IIM), de 2016 jusqu'à 2023.

Jusqu'en 2022, La chaire ICiMa, sous la direction de Raphaèle Fleury et Cyril Thomas, a mené des projets de recherche théorique et appliquée autour de trois axes : les matériaux utilisés pour le spectacle, le geste et le mouvement, la terminologie des arts du cirque et de la marionnette. Ces trois orientations ont donné naissance à des chantiers allant de l'innovation textile—avec des prototypes de combinaisons anti-brûlure pour des acrobates—à la mise en valeur des traces des processus de création ; de l'étude du cycle de vie des matériaux permettant de construire des marionnettes à la notation du mouvement, via l'adaptation du système Benesh à de nouveaux arts.

Cette publication a été mise en œuvre par Esther Friess, Noémie Géron et Gaëtan Rivière. Elle vient rendre compte de l'avancée d'un ces chantiers et des recherches menées plus spécifiquement entre 2020 et 2022.

Articuler archives manuscrites, connaissances marionnettiques et approche linguistique

L'exemple du fonds Chesnais

Stéphane Riou

Le 16 mai 2018 à Charleville-Mézières, lors de la présentation d'un état d'avancement du chantier de recherche « Terminologie multilingue des arts de la marionnette », les étudiant.e.s de la 12e promotion ont pu découvrir la représentation d'un étrange dispositif que Jacques Chesnais avait observé 67 ans plus tôt en Espagne, lors des fêtes d'une rue de Gérone le 25 août 1951 : il s'agissait d'une poupée suspendue au milieu de la rue à une tige de bois, reposant sur deux balcons d'où on la faisait tourner. L'illustration, dessinée, coloriée et légendée de la main même de Jacques Chesnais — rappelons le, élève de Fernand Léger — est issue d'un classeur de 148 pages. Chesnais lui attribuant comme titre « Les comédiens de bois ».

Le fonds Chesnais : un gisement anthropo-linguistique encore inexploité pour la connaissance du monde des arts de la marionnette

Raphaèle Fleury[1] dans son article dénommé " Collectionner pour apprendre " résume parfaitement Jacques Chesnais dans son lien aux arts de la marionnette : c'est un " marionnettiste autodidacte ". Il devient collectionneur pour contourner le secret qui entoure la plupart des grandes familles de marionnettistes " les Piccoli refusent de montrer leurs coulisses, cachent leurs contrôles […] " ou encore " il rapporte en janvier 1951 les circonstances compliquées dans lesquelles Dervaux, beau-frère de Léopold Richard à Roubaix, a fini par lui céder une de ses poupées non sans l'avoir préalablement désensecrétée […]" : "Cette poupée qui est complète, moins le contrôle et les fils, a été remontée par moi et conformément à ce qui était. Malheureusement, en désensecrétant la poupée, M. Dervaux a abîmé les attaches des jambes, le bois étant pourri j'ai du mettre de longs pitons qui ne marchent pas aussi bien que les clous cavaliers qui étaient là ".[2]

Jacques Chesnais (1907-1971)

Jacques Chesnais est un marionnettiste français, auteur de plusieurs ouvrages sur l'histoire des marionnettes. En plus d'être un collectionneur

---

[1] Raphaèle Fleury, 2012, " Collectionner pour apprendre. La collection de Jacques Chesnais ", in *PUCK* n°19, Collections et Collectionneurs, Institut International de la Marionnette, Charleville Mézière.

[2] Chesnais, *Carnet de notes n°2*, 3 janvier 1951, in *ibidem*.

passionné et méthodique des arts de la marionnette, il va, par ses carnets et manuels les documenter. Lors de chacun de ses déplacements, Jacques Chesnais va à la rencontre des marionnettistes locaux. En plus de collectionner ouvrages et marionnettes, il va démonter et dessiner chacun des objets acquis.

Le fonds Chesnais à l'Institut International de la Marionnette

Le fonds Chesnais de l'IIM s'est constitué en 4 versements. Le premier fût fait du vivant de Marion Chesnais, marionnettiste et fille des marionnettistes Madeleine et Jacques Chesnais, qui fit don de la bibliothèque de son père relative à ses recherches sur la marionnette, à l'issue d'une enquête menée pendant 5 ans sur les archives et la collection Chesnais pour le Portail des Arts de la Marionnette.

Un deuxième enrichissement fut fait au moment de la vente aux enchères du 21 juin 2014, par l'acquisition de patrons de costumes et de gaines et de quelques gravures. A son décès survenu le 16 mars 2016, Marion Chesnais a légué à l'Institut International de la Marionnette l'ensemble des documents et objets relatifs aux arts de la marionnette qui restaient à son domicile. Celui-ci est entré à l'Institut en deux nouveaux versements, l'un en octobre 2016 (prélevé à son domicile de la rue d'Assas à Paris), l'autre en décembre 2016 à Bruxelles. Ce legs, constitué d'un ensemble de plus d'une centaine de lots, est extrêmement riche : des marionnettes (par exemple la sublime "Tête du jardinier", datant de la première époque de l'oeuvre de Chesnais, celle de la Branche de Houx – voir photo), une collection de théâtre de papier[3], des correspondances (avec, liste non-limitative, Géza Blattner, Henryk Jurkowski, Guentleur, Albrecht Roser), des comptes-rendus et bulletins d'associations professionnelles, des photographies et également des manuscrits de la main de Jacques Chesnais avec des dessins, des tapuscrits, ainsi qu'une partie de son journal de bord. Une sélection de documents de ce fond établi par Marion Chesnais est numérisée et consultable dans le Portail des Arts de la Marionnette depuis 2014.

Tête de marionnette à gaine du Jardinier pour le numéro éponyme du Théâtre de la Branche de houx, Jacques Chesnais.

[3] Celle-ci devrait faire prochainement l'objet d'une campagne de numérisation pour le chantier Terminologie multilingue de la chaire ICiMa.

Quand l'axe de recherche "Terminologie multilingue des arts de la marionnette" de la chaire ICiMa rencontre le fonds Chesnais

On l'a compris, par sa nature et sa diversité, le fonds Chesnais constitue une ressource particulièrement intéressante pour l'axe "Terminologie multilingue" de la chaire ICiMa : sa numérisation et son traitement vont permettre de faire l'acquisition d'informations (entités nommées, vocabulaire métier, nom et type de marionnettes, couleurs associées) sur le monde des arts de la marionnette entre 1934 et 1971. En effet, le souci du détail, les descriptions et le fait de répertorier dans ses carnets ses rencontres avec les « comédiens de bois » et les professionnels du métier sont de précieuses informations qui permettent de mieux appréhender rétrospectivement le fonctionnement et les échanges des arts de la marionnette de cette époque.

Les cahiers et la compagnie les Comédiens de bois : sélection du corpus

Parmi les éléments du 3e versement du fonds Chesnais datant du 20 octobre 2016, deux ensembles ont prioritairement retenu notre attention en raison de la densité d'information qu'ils contiennent. Il s'agit d'un lot de quatre carnets manuscrits de Jacques Chesnais et de deux grands classeurs annotés "dessins Marionnettes" contenant de nombreux dessins et croquis sur les marionnettes.

Les cahiers : un journal de bord

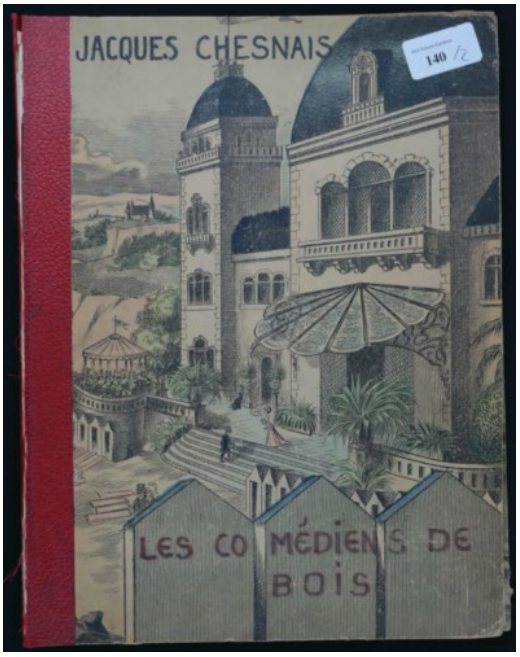
L'ensemble composé de quatre carnets manuscrits est le journal de bord de Jacques Chesnais. Il comprend deux cahiers où sont scrupuleusement consignées les informations relatives aux représentations des Chesnais (avec le Théâtre de la Branche de Houx, puis avec les Comédiens de bois de Jacques Chesnais) 1934 à 1963 et de 1964 à 1969 : on y trouve titres, dates et lieux de représentation, et chacune est numérotée.

Dans les deux autres carnets, Jacques Chesnais a consigné diverses anecdotes et observations au fil de ses lectures, de ses rencontres, de ses voyages, de ses recherches sur la marionnette. L'un porte sur la période 1943-1953 et s'ouvre en octobre 1943 par la présence à Charleroi d'un théâtre de marionnettes à fils portant le nom de théâtre "Farfadet", le second sur la période 1962-1971.

Les comédiens de bois : support visuel et annotations

Le second ensemble, composé de deux grands classeurs de dessins et croquis annotés sur les marionnettes porte la mention “Les Comédiens de bois”[4].

Le classeur représenté ci-contre comporte 148 pages, chacune pouvant contenir un dessin, un plan coté, du texte ou un schéma annoté….



L’écriture manuscrite est un frein à l’accessibilité de l’information. Si l’océrisation[5] de texte dactylographié donne de très bons résultats à l’heure actuelle (encore plus si le texte est généré directement au format numérique), la reconnaissance automatique de l’écriture manuscrite est beaucoup plus problématique. C’est le cas ici où nous partons d’un ensemble de manuscrits (soit un corpus dit “HWT”, abréviation de l’anglais Hand Written Text) hétérogène dont on veut extraire des entités nommées. Plusieurs défis devront être relevés : en premier lieu la reconnaissance du texte manuscrit à proprement parler. Nous avons également une alternance entre les textes et les dessins dont il faut garder le lien, mais également des dessins annotés. Le fait que l’écriture manuscrite bien que contrainte par la matérialité de la page s’affranchît des zones et des espaces textuels classiques (tel ajout de texte va débuter dans l’espace marginale et continuer sur son vis-à-vis ou en note de bas de page au verso[6]). Un autre défi, majeur, est le repérage et la catégorisation d’entités nommées d’un domaine où justement le lexique et les savoirs sont en cours de constitution.

Quelles méthodes utiliser pour la reconnaissance de texte manuscrit (HWT[7]) ?

[4] “Comédiens de bois” était le nom que Chesnais avait donné à sa compagnie (les Comédiens de bois de Jacques Chesnais), mais cette mention en couverture d’un ensemble de croquis d’objets qui ne sont pas seulement ses créations semble indiquer qu’il pouvait utiliser cette expression pour désigner les marionnettes de façon générale.

[5] L’océrisation est l’étape qui lors de la numérisation d’un document permet une conversion en mode texte d’une image. Cette conversion automatique effectuée par un logiciel d’OCR (optical character recognition) demande néanmoins une supervision des résultats du traitement. L’océrisation permet à minima d’identifier les zones de texte et de reconnaître les chaînes de caractères (mots) et s’efforce de faire correspondre sur la copie numérique un positionnement des termes identique au document source permettant ainsi une recherche dite « plein texte ».

[6] Ce qui va justifier l’usage de la pseudo balise —explicitée infra— (X ::).

[7] HWR signifie en anglais Handwritten Word Recognition c’est-à-dire reconnaissance de mot manuscrit qui entre dans l’approche plus globale de Handwritten Text Recognition.

HWT & cloud computing

En terme de cloud computing8, les deux grands acteurs du marché sont Azure (Microsoft) et AWS (Amazon). Ils proposent chacun des services associés au traitement du langage de manière générale et à la reconnaissance d'écriture en particulier. Malheureusement, au moment de l'étude – et à date de publication du présent article – ces services ne sont pas encore déployés pour le français de France (Fr-FR) mais seulement dans le meilleur des cas pour du français du Canada (Fr-CA). Ce qui en a exclu l'usage pour l'analyse des documents Chesnais.

Crowdsourcing : reconnaissance par myriadisation

Le crowdsourcing (également appelé production participative, externalisation ouverte, peuplonomie en fonction du secteur d'activité où il est utilisé) consiste aujourd'hui à faire produire des données par une foule de personnes (de façon rémunérée ou bénévole). Gilles Adda[8] propose, dans une perspective d'annotation de corpus, le terme de « myriadisation » comme néologie française pour crowdsourcing, et de « travail parcellisé » pour celui de *microworking*, opération consistant à segmenter en petites tâches indépendantes et autonomes une tâche complexe. Pour l'exploitation scientifique, ce type de méthode soulève la question de l'expertise des contributeurs (et donc de la validité des données[9]), ce que Karën Fort traite en inversant la conception classique : elle considère que « la myriadisation ne consiste pas à faire produire des données à une foule de non-experts, mais plutôt à identifier des experts de la tâche (en l'occurrence, d'annotation) dans la foule[10] ». L'identification des experts peut se faire soit en mettant en place des procédures de recoupement statistique, de modération humaine a posteriori, pouvant évoluer vers une validation a priori de contributions dont les auteurs ont statistiquement démontré par la qualité de leurs premières contribution leur adhésion à la démarche scientifique établie pour le projet. C'est méthode permettant de sortir du caractère orienté de l'expert unique[11].

Pour la chaire ICiMa et dans ce contexte précis, la reconnaissance par myriadisation consiste à impliquer un groupe ouvert de personnes sur des tâches de travail parcellisé de production ou de contrôle des données dans une

[8] Benoît Sagot, Karen Fort, Gilles Adda, Joseph Mariani, Bernard Lang. Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé. TALN'2011 – Traitement Automatique des Langues Naturelles, Jun 2011, Montpellier, France. inria-00617067

[9] Cette expertise peut jouer à double niveau, comme c'est le cas pour nous : expertise métier (qu'on peut identifier par le fait de pouvoir attester d'une pratique ou d'une compétence donnée, par exemple, le fait d'avoir déjà construit des marionnettes qui ont donné lieu à un spectacle présenté en public) et expertise de contribution, c'est-à-dire la capacité d'une personne à saisir des données dans le respect du protocole garantissant l'exploitabilité scientifique de celles-ci.

[10] Karën Fort, « Experts ou (foule de) non-experts ? la question de l'expertise des annotateurs vue de la myriadisation (crowdsourcing) », Corela [En ligne], HS-21 | 2017, mis en ligne le 20 février 2017. URL : http://journals.openedition.org/corela/4835 ; DOI : https://doi.org/10.4000/corela.4835

[11] Mathieu Lafourcade & Alain JOUBERT (2013) : Bénéfices et limites de l'acquisition lexicale dans l'expérience JeuxDeMots. In Ressources Lexicales : Contenu, construction, utilisation, évaluation, Linguisticae Investigationes, Supplementa 30, John Benjamins, pages 187–216.

perspective de science ouverte. L'expertise dans ce cadre se trouve bien à deux niveaux : l'expertise métier et l'expertise de contribution tout en évitant l'unicité des sources.

Cependant l'approche par transcription participative implique que les documents puissent être publiés en ligne, et donc que leurs droits soient libérés (par la bascule dans le domaine public, ou par décision de publication sous licence ouverte de la part des ayants-droits). Or, dans le cas qui nous occupe, les documents sont encore sous droit, et la légataire avait à multiple reprises fait mention de son désir qu'une partie au moins fasse l'objet d'une publication livresque. Par respect pour ce projet éditorial, nous ne pouvions donc recourir ici à cette méthode participative.

Transcription manuscrite et intelligence artificielle

L'exploitation de corpus manuscrits peut également se faire en recourant à l'intelligence artificielle. Le fer de lance de cette approche est la plateforme Transkribus issue du "Recognition and Enrichment of Archival Documents" (READ Project). Le principe général de ce type d'approche consiste à entraîner dans un premier temps une intelligence artificielle à reconnaître l'écriture cible puis, une fois le système convenablement entraîné, à appliquer celle-ci au corpus que l'on veut reconnaître. Autrement dit, la méthode implique une phase d'investissement de ressources non négligeable avant de pouvoir lancer une reconnaissance automatique sur les textes manuscrits. L'équipe de Transkribus indique que le corpus d'apprentissage (la phase d'apprentissage) doit être au minimum de 15.000 mots.

Cette méthode, très intéressante sur le plan scientifique et technique, n'a pas été retenue à cause de la temporalité inhérente à notre projet et aux coûts de mise en œuvre trop élevés : les temps d'apprentissages, de traitements, d'acquisition des compétences techniques et de formations sur l'outils étaient bien supérieurs à ceux dont nous disposions. De plus, utiliser cette approche pour un corpus dont la taille totale serait proche de la taille nécessaire pour un corpus d'apprentissage ne présentait aucun gain de productivité. L'important était d'obtenir un corpus utilisable. Et une fois que le corpus transcrit sera disponible, il pourra également servir de corpus d'apprentissage dans Transkribus pour une phase exploratoire ultérieure.

Approche par traitement manuel

L'approche qui a été finalement retenue pour traiter un document (le carnet de bord 1943-1953 s'avérera au bout du compte contenir environ 20 000 mots) est celle du traitement manuel lié à un stage[12]. Ce choix n'est pas incompatible avec les autres approches vues supra, soit comme nous l'avons vu par une

[12] Le traitement manuel n'est intrinsèquement pas lié à un stage. Mais dans notre cas précis, la contrainte même du stage (temps disponible, profil) est entrée en ligne de compte dans le choix de la méthode.

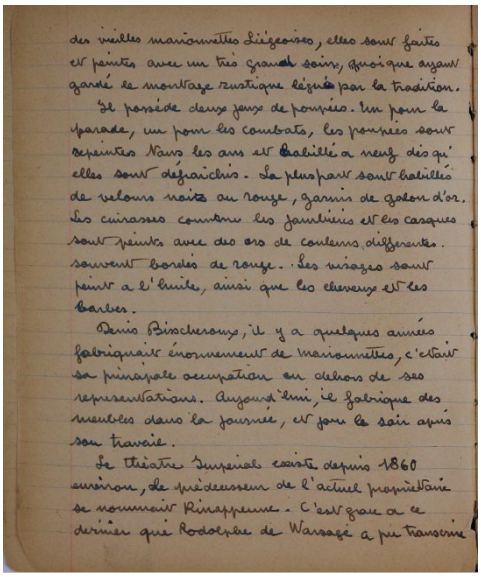
utilisation ultérieure comme corpus d'apprentissage ou comme corpus de contrôle et d'évaluation des contributeurs dans une perspective de myriadisation.

C'est l'approche par traitement manuel que nous allons détailler infra.

Transcrire et baliser manuellement le corpus – Le stage d'Anaïs Britton

Du lundi 13 mai au vendredi 21 juin 2019, le Centre de Recherche et de Documentation de l'Institut International de la Marionnette a accueilli Anaïs Britton, étudiante de l'Université de Caen, pour un stage dans le cadre de l'option "Archives" de son Master Arts, Lettres et Civilisations, parcours "Théâtre".
Les documents avaient préalablement été numérisés à l'IIM à l'aide d'un dispositif utilisant un appareil photo numérique fixé à un support mural éclairé par deux boîtes à lumière.

Installation ayant permis de numériser les archives à transcrire comprenant un appareil photo fixé au mur, deux boites à lumières et un drap noir sur le support horizontal avec repères de placement des documents.
Après la numérisation et la transcription des documents manuscrits, leur encodage se fait aux formats XML et TEI à l'aide des balises <pb> permettant de structurer le document paginé qui a été converti via l'application ABBYY FineReader 14 en PDF « IMAGE SEULEMENT » d'environ 40 Mo pour un total de 152 images[13].


des vieilles marionnettes Liégeoises, elles sont faites et peintes avec un très grand soin, quoique ayant gardé le montage rustique légué par la tradition.
Il possède deux jeux de poupées. Un pour la parade, un pour les combats, les poupées sont repeintes tous les ans et habillé à neuf dès qu'elles sont défraichis. La plusparts sont habillés de velours noirs ou rouge, garnis de galon d'or. Les cuirasses comme les jambières et les casques sont peints avec des ors de couleurs différentes. souvent bordés de rouge. Les visages sont peint a l'huile, ainsi que les cheveux et les barbes.
Denis Bisscheroux, il y a quelques années fabriquait énormement de marionnettes, c'était sa principale occupation en dehors de ses representations. Aujourd'hui, il fabrique des meubles dans la journée, et joue le soir après son travail.
Le théatre Impérial existe depuis 1860 environ, le prédecesseur de l'actuel propriétaire se nommait Rinappeune. C'est grace a ce dernier que Rodolphe de Warsage a pu transcrire

Une page des carnets de bord de Jacques Chesnais et sa transcription. La balise <BP> en tête de capture nous indique que nous sommes à la page 6 du document.

```
<pb facs="Chesnais manuscrit 2016 10 258 C1 - p.0006.pdf"/>
  des vieilles (W::marionnettes Liégeoises), elles sont (W?::faites)
  et (W::peintes) avec un très grand soin, quoique ayant
  gardé le (W::montage) rustique légué par la tradition.
  Il possède deux (W::jeux) de (W::poupées). Un pour la
  parade, un pour les combats, les (W::poupées) sont
  (W::repeintes) tous les ans et (W::habillé) à neuf dès qu'
  elles sont défraichis. La plusparts sont (W::habillés)
  de velours (WC::noirs) ou (WC::rouge), garnis de galon d'(WC::or).
  Les (W::cuirasses) (U::comme) les (W::jambières) et les (W::casques)
  sont (W::peints) avec des (U::ors) de couleurs différentes.
  souvent bordés de (WC::rouge). Les visages sont
  (W::peint) a l'huile, ainsi que les cheveux et les
  barbes.
  (P::Denis Bisscheroux;Bischeroux_Denis), il y a quelques années
  (W::fabriquait) énormement de (W::marionnettes), c'etait
  sa principale occupation en dehors de ses
  (W::representations). Aujourd'hui, il fabrique des
  meubles dans la journée, et (W::joue) le soir après
  son travail.
  Le (I::théatre
Imperial;Marionnettes_Liégeoises_du_théâtre_royal_ancien_impérial_de_Roture)
existe depuis (D::1860)
  environ, le prédecesseur de l'actuel propriétaire
  se nommait (P::?Rinappeune). C'est grace a ce
  dernier que (P::Rodolphe de Warsage;Warsage, Rodolphe de) a pu transcrire


<pb facs="Chesnais manuscrit 2016 10 258 C1 - p.0007.pdf"/>
  la (S::nativité) qui (W::est jouée) tous les ans pendant
  la nuit de Noël.
  En (D::1926). (P::Denis Bisscheroux;Bischeroux_Denis) ayant trouvé
  un acquereur pour tout son (W::materiel), des conservateurs du (I::musee
  de la vie Wallonne), inquiets de voir le dernier
```

Au bas de la capture du document XML, nous voyons le texte du début de la page 7. Entités nommées et extraction d'informations

[13] Les contraintes liées à l'archivage numérique des documents implique la présence de "doublons" dans les prises de vue ce qui se traduit par une correspondance non stricte (une image n'est pas égale à une page unique), ce qui explique que le pdf comprend 152 images pour un document original de 143 pages).

C'est dans le cadre de l'extraction d'informations que les entités nommées sont apparues. Elles sont héritées d'une problématique du traitement automatique du langage, qui succédant à la compréhension globale des textes, se focalisa sur l'extraction d'informations. L'entité nommée est une notion ad-hoc qui, bien qu'efficiente, a fait l'économie d'une véritable épistémologie, notamment linguistique.

Initialement construites autour de 7 types spécifiques regroupés en 3 catégories (ENAMEX, TIMEX, NUMEX) lors des conférences du MUC[14], Maud Ehrmann[15] donne cette répartition issue de MUC-6 en 1996 : PERSON, ORGANIZATION, LOCATION pour ENAMEX (Named Entities Expressions), c'est-à-dire identifier les anthroponymes, les noms d'organisations[16] et les toponymes pour la catégorie originelle des entités nommées ; DATE, TIME pour TIMEX (c'est-à-dire les expressions temporelles des dates et des heures) et MONEY, PERCENT pour les expressions numériques (NUMEX) monétaires et les pourcentages.

En 2002, pour Erik F Tjong Kim Sang[17], elles sont définies comme "Named entities are phrases that contain the names of persons, organizations, locations, times and quantities", qui d'un point de vue de linguistique expérientiel permettrait de les définir comme des syntagmes contenant une référentiation[18] ; Sang y ajoute deux propriétés essentielles, le non-chevauchement (non-overlapping) et la non-récursivité (non-recursive) : une entité nommée ne déborde pas sur une autre (non-overlapping) et elle n'inclue pas, en elle-même, une autre entité nommée (non-recursive) : une entité nommée, dans le cadre de la référenciation, épuise donc sa référence.

Dans leur version "étendue" des entitées nommées Sekine & Nobata[19] en répertorient 200 au sein de trois méta-catégories (qui recoupent celles du MUC-6) : NAME, TIME et NUMEX avec jusqu'à 6 niveaux de profondeurs. Par exemple le type :

>>>>>VERTEBRATE_OTHER FISH REPTILE

Correspond en fait à la hiérarchie :

[14] La série des Message Understanding Conferences, appelées conférences MUC, étaient financées par la Defense Advanced Research Projects Agency des États-Unis.

[15] Maud Ehrmann. Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation. Informatique et langage [cs.CL]. Paris Diderot University, 2008. Français. ⟨tel-01639190⟩

[16] Comprenant les noms d'institutions ou d'entreprises, ce qui correspond globalement à l'onomastique des organisations.

[17] E. F. T. K. SANG. « Introduction to the CoNLL-2002 Shared Task : Language-Independent Named Entity Recognition ». In : Proceeding COL-ING-02 – 6th conference on Natural language learning. Stroudsburg : Association for Computational Linguistics, 2002, p. 1-4.

[18] Pour le dictionnaire Termes et concepts pour l'analyse du discours (2001), la référenciation est l' "[a]cte qui consiste, pour un énonciateur, à désigner un référent à travers l'actualisation d'une séquence linguistique, et résultat de cet acte. Le phénomène de la référenciation met en évidence le rapport entre langage et réalité. Fonction essentielle du langage, la référenciation exploite la dimension puissancielle du signe qui devient, à travers un acte de parole donné, l'instrument d'une dénotation effective".

[19] Sekine, S. et C. Nobata (2004). Definition, dictionaries and tagger for extended named entity hierarchy. In LREC, pp. 1977–1980. Lisbon, Portugal.

NAME>NATURAL_OBJECT>LIVING_THING>ANIMAL>VERTEBRATE>VERTEBRATE_OTHER FISH REPTILE

Sur les 200 types d'entités nommées référencées, 5 composent la méta-catégorie de la temporalité (>TIME_TOP) et 9 pour les expressions numériques (>NUMEX).

À la recherche des entités nommées

Un des objectifs de l'axe terminologique de la chaire ICiMa est le repérage du lexique employé par les artistes. Ceci s'effectue notamment par l'extraction d'entités nommées. Pour cela, 9 des 16 pseudo-balises ajoutées dans la transcription visent directement cet objectif. Elles sont repérées par une parenthèse ouvrante suivi d'une ou plusieurs lettres — ou d'un point d'exclamation "!", caractère permettant au transcripteur d'attirer l'attention du relecteur et surnommé « balise au secours» — suivi de "::" puis du terme ou de l'expression et enfin d'une parenthèse fermante.

Des pseudo balises dans le TEI adaptées à la saisie

Nous avons choisi d'adapter le balisage du texte au format TEI aux impératifs d'une saisie manuelle à l'aide d'un simple environnement de développement intégré (IDE). Pour cela nous avons utilisé les pseudo balises suivantes qui seront dûment converties en TEI V5 s'appuyant sur les préconisations CORLI[20] développées par Christophe Parisse (Modyco, Université Paris Ouest Nanterre) via un programme idoine développé en PHP 7.3.

Les pseudo balises utilisées sur le corpus Chesnais

Le pseudo balisage permet un gain de temps au moment de la saisie et une meilleure accessibilité des informations lorsqu'on n'est pas dans un contexte de traitement XML-TEI. Il s'efforce aussi de prendre en compte le fait que les personnes réalisant l'opération de transcription ne sont pas obligatoirement expertes du domaine.

(C ::) -> Choix = hésitation entre deux termes[21]
(D ::) -> Date
(E ::) -> Mot étranger
(I ::) -> Institutions
(L ::) -> Lieux
(M ::) -> Mesures
(P ::) -> Personnes physiques et personnages
(R ::) -> Rature

[20] CORLI French CLARIN Knowledge Centre for Linguistics of French Language and Beyond. https://corli.huma-num.fr/kcentre
[21] Balise remplacée assez vite par "unclear" (U ::x|x)

(S ::) -> Titres d'œuvres (pièces, spectacles)
(U ::) -> unclear & (U ::x|x) pour indiquer des alternatives
(W ::) -> Vocabulaire métier
(WC ::) -> Couleurs[22]
(WM ::) -> Marionnette-objet
(X ::metaref:texte_comme_il_apparaît;commentaire)
(X :: ref:numéro_de_page_cible;descriptif) & (X :: src:numéro_de_page_source)
(!::n° de page:commentaire) -> commentaires sur le contenu ou la forme du texte

Identités dans le corpus : production d'un fichier tableur

En plus du travail de transcription, Anaïs Britton a effectué un énorme travail de recollection des identités dispersées dans le corpus transcrit, c'est-à-dire dans le carnet de bord numéro 1 et dans le document "Comédiens de bois".

A partir du carnet 1, c'est 471 identités uniques (entitées nommées) qui ont été repérées, réparties entre les catégories "Personnes physiques" (245), "Institution" (58), "Spectacle ou Pièce" (72), "Livre, Revue ou Article" (51) et "Personnage ou Marionnette" (45). Pour le document intitulé "Comédiens de bois", ce sont 100 identités qui ont été extraites ("Personnes physiques" (31), "Institution" (12), "Spectacle ou Pièce" (10), "Livre, Revue ou Article" (2) et "Personnage ou Marionnette" (11)).

Ces données sont intégrables dans le PAM Le Lab' et dans une perspective à long terme dans la base de connaissances JeuxDeMots. À noter que cet exceptionnel et rigoureux travail a fourni des données de références qui pourront servir lors de phases ultérieures d'analyses et d'extractions automatisées.

Bilan

Ce travail a permis d'encoder 148 images[23] extraites des "Comédiens de bois" et 148 autres images extraites du "Carnet de bord" de 1943 à 1953 du fonds Chesnais.

Il a permis de recueillir de précieuses données contenant des identités pour le PAM Le Lab' (https://lelab.artsdelamarionnette.eu) et de disposer de la transcription balisée de manuscrits de Jacques Chesnais qui permet un repérage rapide dans ces documents. Un post traitement rétrocompatible aux normes XML TEI V5 de ces fichiers permettra d'enrichir le corpus examiné dans le cadre du chantier " Terminologie multilingue des arts de la marionnette" de la chaire ICiMa (notes pédagogiques, séries d'interviews de marionnettistes menés lors des deux dernières éditions du Festival Mondial des Théâtres de Marionnettes 2017 et 2019, interviews de constructeurs de marionnettes etc.).

[22] Cette balise spécifique, liée au vocabulaire métier, référence les couleurs que Jacques Chesnais associe dans ses schémas en faisant correspondre le nom d'une couleur à une partie du dessin.

[23] Pour ce travail, une image numérique correspond à une page physique : la totalité des deux manuscrits a bien été encodée soit deux fois 148 pages.

Ce travail, prémices au traitement global du fonds, se veut une démonstration de faisabilité et de l'intérêt de ce que va apporter le traitement terminologique du fonds Chesnais. Il va permettre de rendre accessible la démarche de Jacques Chesnais qui, par son extrême assiduité à dessiner, croquer, prendre des notes, critiquer des spectacles de son œil expert, de même qu'à penser, classifier et typologiser les marionnettes, documente finalement d'une manière extraordinaire le monde des arts de la marionnette. Une telle démarche, pleinement in situ, permet de réintroduire dans nos mémoires et de faire exister 75 ans après, un univers extraordinairement varié et riche que l'on aurait pu oublier.