# Ontologie et TALN : l'anonymisation au service du repérage conceptuel dans le contexte de la SLA

Sonia Cardoso[1], Luis Felipe Melo Mora[2], Marie-Christine Jaulent[2], Xavier Aimé, David Grabli[3], Vincent Meininger[5], Jean Charlet[2,4]

[1] IHU-A-ICM Institut des Neurosciences Translationnelles de Paris,
s.cardoso-ihu@icm-institute.org
[2] INSERM UMRS 1142, LIMICS, F-75006, Paris
Sorbonne Universités, UPMC Univ. Paris 06, UMR_S 1142, LIMICS, F-75006, Paris
Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR_S 1142), F-93430, Villetaneuse
luisfe.melo@gmail.com
[3] Assistance Publique Hôpital Pitié Salpêtrière, Département des maladies du Système Nerveux, Paris
Université Pierre et Marie Curie
david.grabli@psl.aphp.fr
[4]Assistance Publique –Hôpitaux de Paris DRCD, F-75004 PARIS
jean.charlet@upmc.fr
[5]Ramsay General de Santé, Hôpital Peupliers Paris
vincent.meininger@psl.aphp.fr

**Résumé** : L'objectif de notre travail est l'exploitation de la base évènementielle du réseau SLA (Sclérose Latérale Amyotrophique) d'Île-de-France (IDF), pour permettre à long terme, de comprendre les ruptures de parcours de santé. Pour analyser ce corpus une chaine de pré traitement est nécessaire. L'un de ces processus est l'anonymisation, processus consistant à masquer l'ensemble des éléments ne permettant pas l'identification d'une personne. Ce processus de changement de données nominales en catégories sémantiques, permet secondairement une amélioration du repérage des concepts de l'ontologie du domaine, lors de l'utilisation d'outils du traitement automatique de la langue naturelle (TALN).

**Mots-clés** : Ontologie, anonymisation, parcours de soins, sclérose latérale amyotrophique.

## 1 Introduction

L'Ingénierie des Connaissances permet la construction d'ontologies notamment dans le domaine médical qui, associées aux outils de Traitement Automatique de la Langue Naturelle (TALN), permettent d'exploiter des corpus à des fins de compréhension de processus et d'analyse.

L'objectif de notre travail, à long terme, est d'analyser et identifier les indicateurs de ruptures dans le parcours de santé[1] de personnes ayant une pathologie neurodégénérative, en particulier la Sclérose Latérale Amyotrophique (SLA) en exploitant la base de données « évènementielle »

[1] L'Article 14 de la Loi de modernisation de notre système de santé définit dans l'article L. 6327-1 du Code de la santé publique, le parcours de santé complexe, lorsque l'état de santé, le handicap ou la situation sociale du patient rend nécessaire l'intervention de plusieurs catégories de professionnels de santé, sociaux ou médico-sociaux

crée dans le cadre du réseau SLA IDF[2]. Ce réseau est représentatif du suivi de ces patients, puisqu'il accompagne 92% des patients SLA en Île de France (Cordesse *et al.*,2015).

A ce jour la base contient 2245 dossiers patients soit plus de 35 000 évènements. Les événements de coordination sont sous forme de données textuelles non structurées organisées chronologiquement. Ces derniers sont polymorphes dans leurs structures et contenus, ils peuvent contenir des demandes émises par les agents (patients, familles, professionnels), les réponses et actions mises en place par les coordinateurs et des informations de type comptes rendus médicaux.

Pour atteindre l'objectif visé, différentes étapes préliminaires sont nécessaires au traitement des corpus. Nous présenterons les étapes réalisées jusqu'à maintenant. Dans la section 2, nous décrivons le processus de construction d'une ontologie du domaine avec les diverses réflexions menées lors de cette étape pour *(i)* le choix des concepts en lien avec les ontologies et classifications existantes, *(ii)* le choix des labels préférés et synonymes spécifiques aux corpus. La section 3 concernera le processus d'anonymisation et les outils utilisés (plus spécifiquement leur apport dans le repérage d'entités. La section 4 nous permettra de replacer ce travail dans le contexte de l'architecture de traitement mise en œuvre et nous terminerons (section 5) par les limites de ce travail et les perspectives que nous envisageons.

## 2 Méthodologie de construction de l'ontologie

De nombreuses ontologies ont été réalisées dans le domaine de la médecine (comme par exemple *Ontolurgences* (Charlet *et al.*, 2012), *Bilingual Ontology of Alhzeimer's disease and Related Diseases*, ONTOAD (Dramé *et al.*, 2014), ou bien dans le domaine de la coordination des soins infirmiers comme la *Nursing Care Coordination Ontology*[3], NCCO (Popejoy *et al.*, 2014)). Cependant aucune à ce jour ne regroupe les maladies neurodégénératives, la coordination de parcours de santé ou les aspects sociaux et médico-sociaux spécifique au système français.

Nous nous sommes inspirés de la méthode ARCHONTE développée par B. Bachimont (2002) pour construire notre ontologie.

Une première modélisation des actes de coordination fut réalisée en utilisant les diagrammes de séquence ayant « pour but de décrire les modalités de communication entre les objets d'une application, d'un processus ou d'une organisation » (UML 2, 2006). La coordination de parcours de santé consiste en l'interaction d'*agents* (patients, neurologues coordinateurs SLA, …) réalisant des *actions* (demandes, communications…) à destination d'autres *agents* afin d'intervenir sur des *états* ou des *objets* dans des lieux spécifiés (Cardoso *et al.*, 2016).

Le choix et l'agencement des concepts s'est fait en tenant compte de modèles et de classifications existants comme l'ICF (International Classification of Functioning, Disability and Health)[4]. Ces concepts concernent les activités de vie quotidienne, les fonctions de l'organisme ou encore la Classification et terminologie des produits d'assistance pour personnes en situation de handicap (ISO 9999) dans le cadre des aides techniques.

Pour chaque concept, un travail sur deux axes fut mené. Tout d'abord (1) une recherche dans l'ensemble des corpus des synonymes, acronymes, abréviations désignant un même concept. En effet, le coordinateur peut utiliser différentes formes d'un même terme. Par exemple le concept de *Médecin Traitant* est transcrit de huit manières différentes : *méd traitant, MT, MDT, med tt, méd traitant, med ttt* ou encore *médecin de famille*. Afin d'harmoniser la saisie des abréviations et des acronymes utilisés, un travail collaboratif fut mené avec les coordinateurs aboutissant à la création de listes définies. Ces listes sont utilisées secondairement dans le travail d'anonymisation.

[2] http://reseau-sla-idf.fr

[3] http://purl.bioontology.org/ontology/NCCO

[4] http://apps.who.int/classifications/icfbrowser/

Le second axe de travail (2) est l'alignement d'une partie des concepts (travail toujours en cours) avec les classifications et ontologies françaises existantes comme ONTOPYSCHIA (Richard *et al.*, 2013), ONTOLURGENCES (Charlet *et al.*, 2012) et MENELAS (Charlet *et al.*, 2012. Nous avons également utilisé la plateforme HeTOP (Health Terminology / Ontology Portal)[5] qui contient plus de 222 800 définitions (Grosjean *et al.*, 2011) et regroupe un ensemble de terminologies et ontologies du domaine médical et les identifiants Unified Medical Language System (UMLS).

Il nous a semblé nécessaire d'utiliser des classifications et ontologies de référence car notre objectif à long terme est d'appliquer les outils que nous développons à d'autres bases événementielles utilisées pour la coordination notamment la maladie de Parkinson. A ce jour, l'ontologie est constituée de 2946 concepts et nous travaillons à la désignation et à la mise en place des relations entre les concepts.

L'ontologie créée est utilisée secondairement dans le système GATE pour servir de ressource linguistique, permettant le repérage et l'annotation des concepts dans les corpus.

## 3 Travail d'anonymisation

Pour exploiter les corpus il est nécessaire d'anonymiser les données comme le recommande la Commission Nationale Informatique & Libertés (CNIL) et les autorités de protection des données européennes[6] (2014). Ce processus implique de retirer suffisamment d'éléments pour que la personne concernée ne puisse plus être identifiée. Les évènements rédigés sont nominatifs (nom, prénom), temporalisés (date), et localisés (ville, nom de structure). Les informations à anonymiser sont nombreuses. Pour ce travail spécifique, nous avons collaboré avec le LIMSI[7] afin d'utiliser les outils d'annotations qu'ils ont développés : le système d'apprentissage statistique WAPITI[8] (Lavergne *et al.*, 2010) et un système à base de règles et de lexiques DARK (Data Annotation using Rules and Knowledge )[9].

Nous avons défini vingt-deux catégories faisant référence au type sémantique des données à anonymiser dans les corpus regroupées en cinq catégories principales : (1) les *agents* qui concernent des personnes physiques et indiquent leurs « fonctions » *patients, entourage, neurologues* ; (2) les *dates* qui regroupent l'ensemble des éléments temporelles (jour mois années) ; (3) les *structures* pour les *associations de patients*, les *hôpitaux*, les *structures médico-sociales* ; (4) les *identifiants numériques* pour les *numéros de dossiers*, *numéro de téléphone* ; (5) les *lieux* faisant référence à une localisation spatiale : *pays*, *villes*, *départements* et *adresse*.

### 3.1 Outils d'anonymisation utilisés et résultats

A l'image des approches appliquées en fouille de textes, les approches utilisées en désidentification automatique reposent sur deux grandes familles de méthodes : les méthodes à base de règles (généralement implémentées sous la forme d'expressions régulières) et de listes (listes d'entités, dictionnaires, etc..), dites « méthodes symboliques » et les méthodes à base d'apprentissage statistiques reposant sur le repérage des entités à partir d'un corpus annoté (Grouin, 2013). La première étape a consisté à anonymiser manuellement les données de cinquante-quatre dossiers (soit 2311 évènements) extraits de la base de façon aléatoire. Dans un second temps, ces dossiers anonymisés ont servis de référentiel pour le système d'apprentissage WAPITI, afin de créer un modèle d'apprentissage pour annoter

[5] http://www.hetop.eu/hetop/

[6] Groupe de travail « Article 29 » sur la protection des données. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

[7] https://www.limsi.fr/fr/

[8] https://wapiti.limsi.fr

[9] https://perso.limsi.fr/grouin/inalco/1617/index.html

automatiquement de nouveaux corpus. Il est apparu rapidement des difficultés et des erreurs en utilisant le système d'apprentissage seul, du fait des spécificités des corpus non structurés et hétérogènes. En particulier, n'étaient pas détectés correctement le type des personnes (*nom de patient annoté comme professionnel*) et l'oubli de certaines dates et lieux. Pour diminuer ces oublis et erreurs il fut décidé d'utiliser, en complément du système WAPITI (système à base d'apprentissage), un outil à base de règles, DARK, utilisant des lexiques et expressions régulières. Nous avons créé 13 listes d'expressions régulières (nom de patient[10], liste de ville et de pays, …) afin de fournir au système les termes à anonymiser.

Un nouveau corpus constitué de 15 dossiers (soit 431 évènements), tirés de façon aléatoire sur la base, fut testé pour mesurer la performance des deux systèmes en utilisant successivement (1) le système par apprentissage (WAPITI) puis (2) le système à base de règles (DARK). Les résultats obtenus donnèrent un rappel de 0,81 ainsi qu'une précision de 0,90 et une F-mesure de 0,85. Les erreurs obtenues correspondent à des heures annotées comme des dates du fait de leur format.

Initialement nous avions créé, dans le corpus de test, une distinction entre les *patients* et l'*entourage* car, d'un point de vue clinique, il est important de comprendre ce qui émane du patient – ses choix de vie, ses demandes, ses besoins – et ce qui émane de l'entourage (famille, amis). Cependant les outils d'anonymisation utilisés, en particulier l'emploi de lexique (liste des noms de patients) dans le système à base de règles ne permettent pas de faire cette distinction, ces éléments devront être pris en compte lors de l'analyse clinique des parcours.

### 3.2 Utilisation du corpus anonymisé

Une fois le corpus anonymisé, des traitements doivent être effectués pour permettre le repérage des concepts, des informations contextuelles sur ces concepts et des relations existants entre eux. Pour cela nous utilisons des outils développés sur la plateforme GATE[11] pour annoter les corpus, intégrant des applications utilisant l'ontologie du domaine comme ressource lexicale. Les ontologies et les outils de TALN permettent ensuite un travail de recherche intéressant comme le souligne Bodenreider (2008) : « The terminological component of biomedical ontologies is an important resource for natural language processing systems [45] and supports knowledge management tasks such as annotation (or indexing) of resources, information retrieval, access to information and mapping across resources. » et comme cela est confirmé par Charlet *et al.* (2015).

Lors du travail d'anonymisation nous avons émis l'hypothèse, que ce processus aurait un impact, qualitatif et quantitatif, sur le repérage des entités nommées lors de l'utilisation de GATE, en les remplaçant directement par les catégories les subsumant. La transformation d'une donnée nominale (JEAN DUPONT), non défini dans l'ontologie, en donnée identifiant un concept l'*agent* (*Patient*) défini dans l'ontologie permet d'un point de vue conceptuel de repérer les interactions entre les *agents* et les *actions*. Le tableau 1 illustre ces modifications.

| | |
|---|---|
| Corpus original | Appel de Jean DUPONT qui a sollicité Claire MARCHE pour obtenir un certificat médical du Dr MUSCLE pour son dossier MDPH[12], souhaite avoir des nouvelles. |
| Corpus anonymisé | Appel de ***Patient*** qui a sollicité ***Coordinateur SLA*** pour obtenir un certificat médical du ***Neurologue*** pour son dossier MDPH, souhaite avoir des nouvelles. |

***Tableau 1*** - *Exemple d'un événement transformé par l'anonymisation*.

D'un point de vue quantitatif, l'anonymisation permet d'augmenter le nombre de concepts annotés. Nous passons ainsi de 2908 concepts annotés dans le dossier initial à 3578 concepts pour le dossier anonymisé.

[10] Il faut noter que cette liste est évidemment disponible dans le logiciel d'enregistrement des événements et que, dans l'optique d'une mise en œuvre de notre système en routine, elle pourra être remise à jour régulièrement.

[11] https://fr.wikipedia.org/wiki/Architecture_générale_pour_le_traitement_de_texte

[12] Maison Départementale des Personnes Handicapées

D'un point de vue qualitatif, l'anonymisation permet de mieux repérer les « *agents* » acteurs dans le parcours de santé. En effet, la coordination de parcours de santé repose sur l'interaction d'*agents* réalisant des *actions* à destination d'autres *agents* portant sur des *objets*. L'anonymisation permet de comprendre les interactions et les interdépendances des *agents* – et des *structures* – mais aussi de la prépondérance de la présence *versus* absence de certains. En effet, le type de structures sollicitées et le type d'agents impliqués dans la coordination est un facteur important de compréhension des parcours de santé (qui sont les demandeurs, qui sont les agents donnant l'alerte sur la dégradation d'une situation clinique ou familiale comme l'épuisement).

Le processus d'anonymisation va donc avoir une action sur son objectif premier qui est de désidentifier des données afin de pouvoir exploiter des corpus à des fins de recherches, mais aussi améliorer le repérage des concepts par les outils du TALN. Bien que nécessaire et importante dans la démarche d'exploitation des corpus, l'anonymisation ne reste qu'une des étapes du flux de données mis en place dans GATE que nous décrivons dans la section suivante.

## 4 Traitements effectués par GATE

Les textes anonymisés sont ensuite intégrés dans le flux de données de GATE défini comme suit : (1) correction orthographique en utilisant successivement *(i)* un dictionnaire de la langue française (Hunspell), *(ii)* un dictionnaire du domaine (vocabulaire d'UMLS en français) et le lexique de notre ontologie. Une fois un mot mal orthographié repéré, un ensemble de suggestions est proposé par le système en utilisant les termes dénotant les concepts de l'ontologie du domaine et le dictionnaire Hunspell. Nous appliquons, pour ce faire, la distance d'édition Damerau-Levenshtein à chaque mot mal orthographié et ses suggestions, pour déterminer le meilleur choix. Il convient de noter que la distance d'édition ne peut pas être plus grande qu'un seuil spécifié, dans ce cas, la correction ne sera pas faite et le mot restera mal orthographié. (2) reconnaissance de concepts (définis dans l'ontologie) identifiant les extraits du texte qui font référence à ces entités. Pour mener à terme cette tâche, une lexicalisation (PoS : Part of Speech, lemme, etc.) des ressources de l'ontologie (Classes, Instances, Propriétés) est nécessaire. Par la suite, un appariement est effectué entre ces lexicalisations et des fragments textuels. (3) peuplement de l'ontologie à partir des instances repérées dans les corpus.

## 5 Conclusion et perspectives

Si les processus d'anonymisation et de prétraitements réalisés sur les corpus apportent des bénéfices sur l'analyse et l'exploitation des corpus d'un point de vue conceptuel, de nombreuses étapes restent encore à mener. Nous envisageons de travailler prochainement sur la prise en compte des notions hypothétiques (*il est possible que le patient rentre à domicile demain*) et niées (*le patient ne rentrera pas demain à domicile*) présentent dans les corpus, toutes deux ayant un impact sur la compréhension des évènements.

Le second axe de travail sur lequel nous souhaitons avancer est la temporalité. Actuellement, nous avons mis un élément « date » afin d'anonymiser les corpus ; cependant la temporalité est un élément important à prendre en compte lorsque l'on parle de parcours de santé. Un des axes de travail sera de décider d'un mode de représentation formel du temps dans l'ontologie.

Enfin, jusqu'à présent, l'ensemble des outils furent testés pour le corpus du réseau SLA, l'un des objectifs de notre travail est de transposer ces outils à d'autres bases de coordination neurologique. Des essais ont été faits sur la base de la maladie de Parkinson et l'hypothèse de la transposabilité semble valide, même si des éléments spécifiques devront être implémentés.

**Remerciements**

Nous souhaitons remercier Cyril Grouin pour sa participation active, ses outils et ces conseils dans la cadre du travail d'anonymisation.

**Références**

BACHIMONT B., ISAAC A. & TRONCY R. (2002) Semantic Commitment for Designing Ontologies: A Proposal. In A. GOMEZ- PÉREZ & V. BENJAMINS, Eds., 13th International conference on knowledge Engineering and Knowledge Management (EKAW'02), volume 2473 of Lecture Notes in Artificial Intelligence, p.114-121, Sigüenza, Espagne: Springer Verlag.

BONDENREIDER O. (2008) Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. Yearbook of medical informatics, p 67-69

CARDOSO S., AIME X., MELO MORA L-F., JAULENT M-C., GRABLI D., MEININGER V., CHARLET J. (2016) Les ontologies pour aider à comprendre les parcours de santé dans le cadre des maladies neurodégénératives. Conférence: IA & Santé 2016 - Deuxième Atelier sur l'Intelligence Artificielle et la Santé, At Montpellier

CHARLET J., DECLERCK G., DHOMBRES F., GAYET P., MIROUX P.ET VANDENBUSSCHE P.-Y. (2012) Construire une ontologie médicale pour la recherche d'information : problématiques terminologiques et de modélisation. In : Szulman S., coordinateur. Actes des 23es Journées Ingénierie des Connaissances, Paris, France, 27-29 juin, p. 33-48.

CHARLET J., BACHIMONT B., MAZUEL L., DHOMBRES F., JAULENT M. ET BOUAUD J. (2012) OntoMenelas : motivation et retour d'expérience sur l'élaboration d'une ontologie noyau de la médecine. Technique et Science Informatiques, 31(1).

CHARLET J., DARMONI S -J. (2015) Knowledge Representation and Management. From Ontology to Annotation. Findings from the Yearbook 2015 Section on Knowledge Representation and Management. Yearbook of Medical Informatics, p 134-136.

CORDESSE V., SIDOROCK F., SCHIMMEL P., HOLSTEIN J., MEININGER V. (2015) Coordinated care affects hospitlization and prognosis in amyotrophic lateral sclerosis: a cohort study. BMC Health Services Research

DRAME K., DIALLO G., DELVA F., DARTIGUES J-F., MOUILLET E., SALAMON R., MOUGIN F. (2014) Reuse of termino-ontological ressources and text corpora for building a multilingual domain ontology: An application to alzheimer's disease. Journal of Biomedical Informatics 48, 171-182

GROUIN C. (2013) Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique, PhD Thèse. Université Pierre et Marie Curie, Paris, France.

GROSJEAN, J; MERABTI, T; DAHAMNA, B; KERGOURLAY, I; THIRION B; SOUALMIA LF & DARMONI, SJ. (2011) Health Multi-Terminology Portal: a semantics added-value for patient safety. Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety, Studies in Health Technology and Informatics, Volume 166, Pages 129-138.

GOMEZ-PEREZ, A., FERNANDEZ M. AND DE VICENTE A.J. (1996) "Towards a Method to Conceptualize Domain Ontologies", ECAI-96 Workshop on Ontological Engineering, Budapest.

LAVERGNE T., CAPPE O., FRANCOIS Y. (2010) Pratical VeryLarge Scale crfs. Proceedings the 48th Annual Meeting of the Association for Computational Linguistics, 504-513. Uppsal, Sweden

Popoejoy LL., Khalilia M-A., Popescu M., Galambos C., Lyons V., Rantz M., Hicks L., Stetzer F., (2014) Quantifying care coordination using natural language processing and domain-specific ontology. Journal of the American Medical Informatics Association, Volume 22, p 93-103.

PILONE D., PITMAN N. (2006) UML 2 en concentré. Edition O'Reilly, Paris.

RICHARD M., AIME X., KREBS M.-O. & CHARLET J. (2013) Au-delà du DSM : les ontologies comme aide aux classifications descriptives psychiatriques ? 2e édition du Symposium sur l'Ingénierie de l'Information Médicale, Jul 2013, Lille, France.