Une procédure d'anonymisation à deux niveaux pour créer un corpus de comptes rendus hospitaliers

Cyril Grouin[1], Arnaud Rosier[2], Olivier Dameron[2], Pierre Zweigenbaum[1]

[1]LIMSI, CNRS, F-91403 Orsay, France
[2]Inserm U936, F-35000 Rennes, France

Abstract

De-identification is a growing need in medical informatics, and has therefore recently been the subject of renewed interest. De-identification needs to be tuned to the local documents and their specificities, which requires language engineers to work on non-de-identified text. To lower the issues linked to such a situation, we propose a de-identification method which proceeds in two steps. We report experiments on the adaptation of an American de-identifier to French and on the development of a new de-identifier for French patient reports. The latter, evaluated on a set of 23 randomly selected texts, obtains 85 % recall and 91 % precision.

Keywords. Natural Language Processing, Anonymization;

1 Introduction

Dans les travaux qui portent sur l'analyse automatique de textes en langue naturelle, les corpus de textes sont un matériau fondamental. Cette observation générale s'applique également au traitement automatique de textes du domaine médical ou biomédical, qu'il s'agisse d'indexation et de codage automatique [1], de catégorisation de patients[1] ou de repérage de gènes, de protéines et de leurs interactions [2]. Un corpus de textes est utile pour étudier les problèmes à traiter et mettre au point le système d'analyse. Un corpus de textes dans lequel on a ajouté des annotations qui indiquent les résultats attendus [2] permet de plus d'évaluer automatiquement les résultats d'un programme en cours de développement. Il permet enfin, s'il est suffisamment grand, d'entraîner automatiquement un système fondé sur des mécanismes d'apprentissage.

Le travail sur des textes cliniques pose cependant un problème de taille : ces textes ne peuvent être utilisés en dehors du soin des patients que si toutes les marques permettant d'identifier le patient ont été supprimées. Ils doivent donc être *anonymisés* avant d'être mis entre les mains de chercheurs et de développeurs de méthodes de traitement automatique des langues, ou simplement pour être inclus dans une publication (par exemple, étude de cas). Cela crée une contrainte forte qui cause une extrême rareté des corpus disponibles dans le domaine clinique, contrairement par exemple aux corpus de résumés d'articles scientifiques qui sont utilisés dans le domaine biomédical (génomique). Zweigenbaum [3] cite cette contrainte comme l'un des facteurs clés dans le différentiel de développement des

[1] Voir par exemple la catégorisation du statut d'obésité d'un patient (https://www.i2b2.org/NLP/).

travaux de traitement automatique des langues en informatique médicale par rapport à ceux effectués en bioinformatique. Ces besoins ont motivé des travaux sur l'anonymisation automatique de comptes rendus cliniques, que nous décrivons plus bas.

Après une brève revue de travaux proches (section 2) et de notre matériel de départ (section 3), nous présentons ici un travail d'anonymisation en trois points (section 4). D'une part, le travail de mise au point d'une méthode d'anonymisation nécessite lui-même un corpus de développement dont la disponibilité hors d'un service hospitalier pose elle-même le problème de son anonymisation. Nous réduisons cette contrainte en effectuant le plus gros de l'anonymisation à la source, dans le service hospitalier. D'autre part, nous rapportons une expérience de réutilisation d'un anonymiseur réalisé aux États-Unis pour traiter les informations identifiantes [4], dont nous avons converti en français les termes liés à la langue anglaise. Enfin, nous avons développé un anonymiseur directement conçu pour traiter des textes médicaux français, que nous évaluons sur un échantillon de documents de notre corpus. Nous présentons les résultats de son évaluation (section 5), les discutons et concluons sur quelques perspectives (section 6)[2].

2 Contexte et état de l'art

Des travaux sur l'anonymisation automatique de comptes rendus cliniques ont été menés dans différents contextes [5, 6]. Ils ont été récemment dynamisés aux États-Unis avec la problématique des « données de santé protégées » (*Protected Health Information*, PHI). Les États-Unis se sont en effet dotés d'une liste officielle de 18 types d'identifiants dont la suppression dans un document contenant des PHI rend possible la levée des contraintes pesant sur la transmission de ces données et donc de ce document (par exemple à des fins de recherche)[3] : noms et prénoms, toute indication de lieux, dates, âge si supérieur à 90 ans, numéros de téléphone et de télécopie, adresses de messageries électroniques, numéros de Sécurité Sociale, numéro d'enregistrement médical, numéro de complémentaire santé, numéro de compte, numéro de cartes d'identité et de permis de conduire, références sur le véhicule du patient, URLs, adresses IP, numéro de série ou identifiant d'appareil implanté (pacemaker), identifiants biométriques, et n'importe quel autre code ou caractéristique permettant d'identifier un patient tel qu'une photographie du visage, de cicatrices ou de tatouages spécifiques).

Cette liste a facilité le développement de travaux d'anonymisation aux États-Unis. Il n'existe en revanche pas à notre connaissance en France de critères clairs définissant les informations à supprimer d'un texte pour qu'il soit considéré comme anonymisé. En l'absence de ces critères, nous reprenons les principaux parmi ceux employés aux États-Unis, qui sont par ailleurs en cohérence avec des critères retenus après discussion avec la CNIL dans un projet passé [7] : noms, prénoms, lieux, dates et âges.

L'intérêt du partage de données médicales est par ailleurs fortement soutenu par diverses instances publiques, à l'image de l'Institut Canadien de Recherches Médicales (CIHR) qui rend obligatoire, depuis le 1er janvier 2008, la mise à disposition des données utilisées dans le cadre de publications [8]. A cet effet, l'anonymisation des informations médicales personnelles n'en est que plus pertinent.

Une compétition portant sur l'anonymisation de 220 comptes rendus d'hospitalisation a été organisée en 2007 par le projet i2b2[4] [9]. La quasi-totalité des systèmes était fondée sur des

[2] Ce travail a été partiellement financé par le projet Akenaton (ANR-07-TECSAN-001-06).

[3] Office for Civil Rights — HIPAA (http://www.hhs.gov/ocr/hipaa/understanding/summary/) (site visité le 21/1/09).

[4] Informatics for Integrating Biology & the Bedside, http://www.i2b2.org/ (site visité le 21/11/08).

méthodes de catégorisation automatique (SVM, arbre de décision, etc.) ou d'étiquetage de séquence (HMM, CRF), travaillant sur des textes généralement pré-analysés à l'aide d'outils de traitement automatique des langues (étiquetage morpho-syntaxique, etc). Les meilleurs systèmes ont obtenu une F-mesure supérieure à 98 %, avec une précision un peu supérieure et un rappel un peu inférieur[5].

Neamatullah *et al.* [4] remarquent que ces résultats ont été obtenus par apprentissage sur un grand corpus étiqueté à la main au préalable, et qu'il n'est pas évident que cette approche soit transportable à d'autres types de textes. Leur logiciel DE-ID (pour « de-identification » ; voir section 3.2) repère les entités à anonymiser à l'aide de dictionnaires et d'expressions régulières. Évalué sur un corpus de notes infirmières, considéré comme plus difficile que les comptes rendus d'hospitalisation d'i2b2, il obtient un rappel de 96,7 % et une précision de 74,9 %.

Friedlin et McDonald [10] s'attaquent à des sources de textes plus variées provenant de messages HL7 transmis au Regenstrief Medical Record System. Leur système, MEDS[6], emploie également des listes et des expressions régulières. Son évaluation finale sur 7 193 comptes rendus d'anatomopathologie a obtenu un rappel de 99,47 % sur les noms propres et de 96,93 % sur les autres éléments (sa précision sur 2 400 documents variés était de 92 %). Les auteurs affirment que les résultats obtenus par leur système sont similaires à ceux des meilleurs systèmes de la campagne i2b2.

Les listes de MEDS sont constituées à partir de l'UMLS pour les noms propres et noms d'hôpitaux. Les expressions régulières sont employées pour les entités numériques. Ses auteurs ont également considéré qu'il était possible d'améliorer l'anonymisation en ayant recourt à l'étiqueteur/lemmatiseur de MetaMap [11], traitement qui se révèle efficace pour traiter les mots qui appartiennent aux listes de noms propres mais qui, dans le contexte où ils apparaissent, ne sont pas des noms propres (notamment dans le cas d'un mot écrit en minuscules et précédé d'un article ou d'un adjectif : *I examined the pat*, où *pat* désigne le patient mais pourrait être pris pour le prénom *Pat*). Enfin, un dernier module étudie le voisinage des noms propres connus de manière à traiter les éventuelles fautes de frappe.

L'anonymiseur le plus connu pour le français [5] n'étant pas de source libre, donc pas adaptable par un tiers, nous nous sommes orientés vers un anonymiseur disponible au téléchargement, en l'occurrence DE-ID. Il était de plus intéressant d'examiner dans quelle mesure un anonymiseur conçu pour l'anglais était adaptable au français. Lorsqu'ensuite nous avons été amenés à concevoir notre propre anonymiseur, nous nous sommes inspirés de DE-ID et de MEDS.

3 Matériel

Nous décrivons ici nos données initiales : un corpus de comptes rendus collecté pour le projet Akenaton et l'anonymiseur DE-ID [4] introduit plus haut.

3.1 Corpus de comptes rendus

L'objectif étant de constituer le corpus le plus large possible, nous avons décidé de cibler un ensemble varié d'unités fonctionnelles (UF) appartenant au centre cardio-pneumologique d'un centre hospitalier régional universitaire. Ces UF comprennent des

[5] Le *rappel*, ou sensibilité, est la proportion des entités à anonymiser effectivement repérées par le système ; et la *précision*, ou valeur prédictive positive, est la proportion des entités repérées par le système qui devaient effectivement être anonymisées. La *F-mesure* est la moyenne harmonique du rappel et de la précision.

[6] MEDS : Medical Deidentification System.

services d'hospitalisation, des hôpitaux de jour, des services de consultation, une unité d'accueil d'urgence spécialisée ainsi que des services médico-techniques et chirurgicaux, et couvrent donc toute une gamme de prise en charge diagnostique et thérapeutique ainsi que des modes de suivi de patients variés. Trois spécialités étaient concernées : la cardiologie, la chirurgie cardiaque et la pneumologie. Ces spécialités traitent de pathologies thoraciques assez souvent liées sur le plan médical, ce qui signifie une cohérence relative des problèmes médicaux principaux dont il est question dans les comptes rendus créés.

Pour l'ensemble de ces unités, l'intégralité des comptes rendus (hospitalisation et opératoires) et courriers créés entre 2002 et début 2008 ont été extraits du système d'information hospitalier, à l'aide d'une requête de la base de données correspondante. Pour chacun des documents, un fichier contenant le texte a été généré, ainsi qu'un fichier associé, au format XML, qui contient sous forme structurée les métadonnées du système d'information hospitalier au sujet du patient : données d'identification et données de provenance du document. Ces métadonnées seront utiles pour l'anonymisation à la source de ces documents (voir la section 4.1), sous le contrôle coordonné de médecins du centre cardio-pneumologique et du département d'informatique médicale.

3.2 L'anonymiseur DE-ID

L'anonymiseur DE-ID [4], réalisé par plusieurs équipes du MIT[7], se compose d'un ensemble de scripts Perl. Les types d'informations anonymisées par DE-ID reposent sur les 18 catégories spécifiques d'informations définies par l'HIPAA.

Les méthodes utilisées par DE-ID reposent sur une combinaison de moyens. En premier lieu, le logiciel mobilise des ressources linguistiques de deux types : des *dictionnaires* (noms communs et expressions médicales) et des *listes d'entités nommées* (noms, prénoms, hôpitaux, villes, abréviations d'États américains, indicatifs téléphoniques, pays, gentilés, etc). Pour un même type d'entité nommée, une distinction est faite entre les entités ambiguës et les entités non ambiguës (à ce titre, le prénom féminin « France » figurera dans la listes des prénoms féminins ambigus, puisqu'il s'agit également d'un nom de pays).

En second lieu, le logiciel exploite des *listes de déclencheurs* qui permettent de détecter des informations à anonymiser dans le voisinage – les quelques mots qui précèdent et qui suivent – de ces déclencheurs. Ces déclencheurs sont typés pour chaque catégorie d'entité nommée : une liste de particules (*De, Mc, Van…)* et de titres (*Mister, Doctor, Professor…*) pour les noms de famille, des éléments précédant un lieu (*lives in, resident of, comes from*), etc.

Enfin, le logiciel utilise un ensemble d'*expressions régulières* reposant sur une précédente identification de termes du voisinage (un mot sera anonymisé comme nom de famille s'il ne figure pas dans le dictionnaire des noms communs et que le mot qui le précède a été identifié comme un prénom de manière sûre).

En dehors de la trace habituelle des anonymisations réalisées, DE-ID attribue un identifiant numérique unique à chaque nom rencontré dans un texte et affiche cet identifiant dans la version anonymisée. Ceci permet ainsi de conserver une information de co-référence – le suivi de la réapparition des différents noms en corpus – sans sacrifier à l'impératif de l'anonymisation des informations personnelles.

Un autre aspect de la conservation des informations contenues dans les textes tout en

[7] Laboratory for Computational Physiology, Harvard-MIT division of Health Sciences & Technology, Computer Science and Artificial Intelligence Laboratory

respectant le devoir de protection des données consiste à associer à chaque document un « décalage dans le temps » distinct à chaque document mais qui permet de conserver la vraisemblance des intervalles de temps exprimés dans les différentes dates d'un document.

Les résultats obtenus par DE-ID sur le corpus d'évaluation de Neamatullah *et al.* [4][8] varient selon le type d'entité anonymisé : l'anonymisation des noms de cliniciens obtient ainsi un rappel de 99,5 % et une précision de 72,5 % contre un rappel de 76,1 % et une précision de 71,3 % pour les dates avec années. Globalement, le système DE-ID obtient un rappel de 96,7 % et une précision de 74,9 %. Les auteurs mettent également en avant un taux de faux positifs assez faible (de l'ordre de 19,72 faux positifs pour 100 000 mots), avec un taux de faux positifs étrangement plus élevé pour les dates (7,769 pour 100 000 mots) que pour les noms (1,494 pour 100 000 mots).

4 Méthodes

4.1 Anonymisation à la source

Comment mettre au point des méthodes d'anonymisation poussée sur des textes réels sans pour autant travailler sur des données fortement identifiantes ? Nous n'avons pas trouvé de mention de ce problème dans la littérature sur l'anonymisation de textes cliniques. La solution que nous avons mise en place consiste à effectuer une première passe d'anonymisation reposant sur des méthodes simples, qui peuvent être mises en œuvre à la source lors de la collecte des textes, mais permettant de supprimer la quasi-totalité des informations identifiantes fondamentales : le nom, le prénom et la date de naissance du patient.

Ces informations figurent en effet dans la partie structurée du dossier patient, et ont été extraites dans le fichier de métadonnées associé à chaque document textuel. Ces métadonnées contiennent en tout :

- données patient : nom, nom marital, prénom, date de naissance et sexe du patient ;
- données de provenance du document : identifiant de l'UF, date du compte rendu, nom de l'auteur.

L'algorithme d'anonymisation à la source a été conçu pour réaliser les opérations suivantes, au sein de l'entrepôt de données hospitalier :

- remplacer le texte mentionnant le nom du patient (quelle que soit la casse) par une balise <Nom patient> ;
- remplacer de façon similaire le texte mentionnant le nom marital et le prénom du patient ;
- remplacer la date de naissance par une balise <Date naissance patient>, d'après une identification basée sur trois types de motifs : 01/01/70, 01/01/1970 et 01 Janvier 1970.

Pour chaque document, le nombre de remplacements a été comptabilisé et les documents pour lesquels aucune altération du nom ou du nom marital n'était réalisée ont été identifiés et vérifiés manuellement.

Enfin, dans les fichiers de métadonnées, les noms, prénom et date de naissance du patient et le nom du médecin ont été chiffrés selon l'algorithme SHA-256, et les autres

[8] Ce corpus comprend un ensemble de notes de soins infirmiers anonymisées qui ont été réidentifiées de manière plausible ; ce corpus est utilisé pour le développement et l'évaluation de DE-ID.

informations supprimées. Ces fichiers, fournis avec le corpus, conservent donc une information permettant de mettre en évidence les documents parlant d'un même patient, mais sans possibilité de remonter à l'identité du patient. Cette méthode est similaire à celle employée par Quantin *et al.* [12] pour créer des identifiants familiaux chiffrés.

L'ensemble des documents (texte + métadonnées résiduelles codées) constitue le corpus à l'issue de cette première passe d'anonymisation à la source. Chaque texte va ensuite faire l'objet de traitements plus poussés pour prolonger son anonymisation.

Un échantillon de 23 textes a été aléatoirement extrait du corpus à des fins d'évaluation. Chacun de ces textes a été manuellement anonymisé pour servir de référence (*gold standard*) : les expressions de type nom, prénom, date et ville y sont respectivement remplacées par les balises XML <nom />, <prenom />, <date /> et <ville />.

4.2 Conversion partielle en français de l'anonymiseur DE-ID

Devant la qualité des résultats obtenus par le logiciel DE-ID et du fait de sa distribution sous licence GNU autorisant librement sa réutilisation, nous avons essayé de l'adapter au français. Son adaptation à une nouvelle langue passe par plusieurs modifications de difficultés inégales portant sur chacune des trois ressources qu'il utilise.

4.2.1 Dictionnaires et listes

La qualité des anonymisations dépendant pour une part non négligeable de la richesse des dictionnaires et listes utilisées, une attention particulière doit être accordée aux ressources linguistiques utilisées.

Nous avons d'une part employé des lexiques disponibles sur le site de l'Association des Bibliophiles Universels (ABU), une association de bénévoles qui proposent en libre téléchargement depuis son site Internet[9] des listes d'entités nommées et de dictionnaires. Toutes les listes proposées ne sont pour autant pas d'égale qualité ; si le dictionnaire des noms communs se révèle riche et varié (plus de 251 000 formes différentes), la liste des villes françaises est désaccentuée, tout comme la liste des prénoms qui, de plus, intègre une majorité de prénoms anglo-saxons. Il s'agit cependant de ressources utiles qu'il est possible d'adapter (en tentant une réaccentuation automatique), ou pour lesquelles il est peut être envisagé de contraindre les outils de traitement automatique des langues à s'adapter (en étant insensibles aux caractères accentués).

Nous avons par ailleurs complété cet ensemble de listes par nos propres ressources, soit en reprenant des listes précédemment constituées (liste de 33 380 noms de villes), soit en créant de nouvelles ressources (liste de 2 526 noms d'hôpitaux, cliniques, centres hospitaliers français, constituée à partir d'un annuaire de santé sur Internet[10]).

4.2.2 Listes de déclencheurs

Adapter les listes de déclencheurs au français s'est révélé être la tâche la plus facile dans le sens où il s'agit d'une traduction des termes avec une éventuelle complétion des listes utilisées[11]. Le logiciel DE-ID reposant sur des listes de déclencheurs distinctes selon que le déclencheur est attendu avant ou après le mot à anonymiser, la contrainte de l'ordre des mots qui diffère en anglais et en français a ainsi pu être aisément contournée.

---

9 http://abu.cnam.fr/(site visité le 21/11/08).

10 http://www.sanitaire-social.com/(site visité le 14/02/08).

11 En raison de l'ordre des mots différents en français et en anglais, certains déclencheurs présents dans la liste des suffixes en anglais ont été déplacés dans la liste des préfixes en français : les suffixes anglais *Street, Road, Blvd* deviennent les préfixes *Rue, Route, Bld* en français.

4.2.3 Expressions régulières

Nous avons constaté que modifier les expressions régulières constituait la tâche la plus difficile dans ce travail d'adaptation du logiciel au français. En effet, cette adaptation dépend d'une part de la manière dont le logiciel a été conçu (du point de vue de la manière de coder), et d'autre part de l'objectif visé (en l'occurrence, convertir le logiciel pour une utilisation en français, sur un corpus de comptes rendus médicaux).

Si l'adaptation des expressions régulières pour les entités de type numérique (numéros de Sécurité Sociale, téléphone, dates, etc.) ne pose aucun problème particulier, il en est tout autrement des détections d'entités nommées. La réalisation d'expressions régulières étant fortement contextuelle, il ne nous a guère été possible d'adapter les expressions existantes sans les reprendre à la base, ce qui est très chronophage.

4.3 Conception d'un nouvel anonymiseur : Medina

Devant les difficultés rencontrées pour adapter le logiciel DE-ID au français, nous avons fait le choix de développer notre propre anonymiseur pour le français : Medina (Medical Information Anonymization). Nous nous sommes focalisés pour commencer sur les anonymisations de noms, prénoms, dates et villes.

Compte-tenu de l'expérience acquise par les auteurs des logiciels DE-ID et MEDS, nous avons fait le choix de combiner plusieurs méthodes : les expressions régulières pour les entités numériques, les dictionnaires et listes pour les entités nommées, ainsi qu'une seconde anonymisation fondée sur le voisinage des termes déjà anonymisés. Dans la mesure du possible, nous avons essayé de distinguer le programme des ressources utilisées (les ressources sont déportées dans des fichiers distincts). Nous avons également cherché à simplifier la configuration du logiciel en utilisant un fichier de configuration externe au programme.

Nous avons repris les listes et dictionnaires déjà réalisés précédemment lors de la tentative de francisation du logiciel DE-ID. Nous nous sommes par ailleurs inspirés des distinctions faites entre listes ambiguës et non-ambiguës de DE-ID pour nettoyer les listes de noms, prénoms et pays. A cet effet, nous avons extrait de ces trois listes tous les termes qui existent dans le dictionnaire de noms communs[12], de manière à réduire le nombre de suranonymisations futures. Les noms ainsi supprimés de ces différentes listes — en particulier pour les prénoms — se révèlent d'autant moins ambigus que les listes utilisées contenaient une majorité de noms anglo-saxons qui semblent assez étranges à porter en France (*Agace, Dragon, Masculine, Travers,* etc).

Ce programme a été réalisé pour un type de corpus particulier et entraîné sur ce corpus. Parmi les limites de cet exercice, nous pouvons mettre en évidence le fait que nous n'avons utilisé les données médicales provenant d'un seul hôpital. Nous devrions toutefois être en mesure d'obtenir prochainement des comptes rendus opératoires en provenance d'un second hôpital, de telle sorte que nous pourrons tester la généricité du programme sur plusieurs corpus.

4.4 Mode d'évaluation

L'évaluation est réalisée par comparaison des balises attendues avec celles effectivement produites. Il s'agit ici de donner une représentation robuste aux décalages de l'occurrence mais aussi de la position de chaque balise d'anonymisation insérée dans un texte. Pour

[12] Nous avons réduit la liste de noms de familles de 3,47 % à 12 994 noms, celles des prénoms l'a été de 5,4 % à 11 746 prénoms, enfin, celle des noms de villes l'a été de 2,8 % à 32 439 noms.

cela, nous relevons dans le fichier de référence chaque balise avec son contexte immédiat gauche et droit de manière à inscrire la balise dans un contexte discriminant. Nous avons limité le contexte aux quatre caractères qui précèdent et suivent la balise car ils suffisent à localiser la séquence dans le texte[13]. Dans la phrase « *Je vous laisse le soin de demander à Monsieur <nom /> de bien vouloir prendre un rendez-vous* », nous relèverons ainsi la séquence « *eur <nom /> de* », ce qui nous permet de localiser la balise <nom /> dans le texte avec son contexte gauche et droit. A l'inverse, prendre plus de caractères contextuels n'a pas permis de mettre en évidence des cas d'ambiguïté. En revanche, si nous réduisons le contexte aux deux caractères qui précèdent et qui suivent, des cas de séquences ambiguës apparaissent[14] — augmentant artificiellement le nombre de balisages corrects — et l'évaluation s'en trouve alors faussée.

Nous relevons également les balises dans le fichier de résultats de l'anonymiseur. Puis nous avons comparé les balises produites avec celles de la référence et avons évalué les résultats produits en termes de rappel (nombre de balisages corrects rapporté au nombre de balisages attendus) et de précision (nombre de balisages corrects rapporté au nombre de balisages produits).

Parmi les limites de ce type d'évaluation, nous pouvons distinguer deux cas. En premier lieu, cette évaluation repose sur une identité stricte entre balises. Dans le cas d'un nom composé, si la référence substitue le nom composé par une balise alors que le système d'anonymisation retourne deux balises (une pour chaque nom), l'évaluation considérera qu'il y a échec du balisage quand bien même l'anonymisation a été produite. La seconde limite se rapporte au typage de l'information anonymisée. Pour une information donnée (typiquement un nom de famille qui relève d'un prénom[15]), la référence typera cette information en qualité de nom alors que l'anonymiseur pourra la typer comme un prénom. Face à cette différence de typage de l'information, l'évaluation considérera qu'il y a échec du balisage alors même que l'anonymisation aura été produite.

5 Résultats

Le corpus constitué comprend 21 749 documents uniques, concernant 11 964 patients individuels. Comme indiqué plus haut, un échantillon de 23 textes a servi à nos évaluations. Nous distinguons dans le tableau 1 l'évaluation de premier niveau (celle réalisée pour pouvoir sortir ces données de leur hôpital d'origine) de celles de second niveau (DE-ID et Medina), ces dernières poursuivant l'anonymisation sur les fichiers semi-anonymisés de premier niveau.

Tableau 1 : Comparaison des balisages des différents outils d'anonymisation.

	Rappel	**Précision**	**Corrects**	**Ramenés**	**Attendus**
Premier niveau	0,91	1,00	73	73	80
DE-ID francisé	0,63	0,25	103	402	163
Medina	0,85	0,91	139	152	163

[13] Trois caractères semblent suffire mais pour plus de sûreté, nous avons porté à quatre caractères ce contexte.

[14] Les deux caractères renvoient par exemple à la dernière lettre du mot précédent, suivie de l'espace, ce qui se révèle trop imprécis donc ambigu.

[15] Laurent, Martin, etc.

L'anonymisation de premier niveau obtient une excellente précision et un bon rappel. Notre francisation de DE-ID obtient un rappel moyen et une précision très faible. Le programme se déclenche dans un grand nombre de cas, ce qui cause mécaniquement un bruit important (suranonymisation). Medina obtient de meilleurs résultats que la version francisée de DE-ID avec de bons taux de rappel et précision.

A titre d'illustration du fonctionnement de Medina et des types d'erreurs commises, le tableau 2 présente quelques exemples d'anonymisations réalisées. Les expressions marquées *Entrée* sont le résultat de l'anonymisation de premier niveau, celles marquées *Sortie* sont le résultat de l'application de Medina sur cet état initial.

Tableau 2 : Exemples de résultats de Medina. Note : les informations identifiantes de ces exemples ont été modifiées manuellement avant d'être copiées dans le tableau.

Étape	Énoncé
Anonymisation réussie	
Entrée	*J'ai examiné en consultation Madame <Nom marital patient> Michèle, née le 13.1.1943, âgée de 62 ans, pour le contrôle annuel de son stimulateur double chambre.*
Sortie	*J'ai examiné en consultation Madame <nom /> <prenom />, née le <date />, âgée de <age />, pour le contrôle annuel de son stimulateur double chambre.*
Sur-anonymisation	
Entrée	*(le pace maker est actuellement réglé en VVI à une fréquence de 52/mn).*
Sortie	*(le <prenom /> maker est actuellement réglé en VVI à une fréquence de 52/mn).*
Sur- (*Pace Maker*) et sous- (*PECRESSE*) anonymisation	
Entrée	*Le Pace Maker avait été contrôlé au mois de janvier par le PR PECRESSE et ne montrait pas de signe d'usure.*
Sortie	*Le <prenom /> <nom /> avait été contrôlé au mois de <date /> par le PR PECRESSE et ne montrait pas de signe d'usure.*
Autres anonymisations réussies, hors évaluation	
Entrée	*un épisode d'IVG justifiant d'un traitement par Lasilix et la réduction de la posologie de Soprol à 1/jour.*
Sortie	*un épisode d'IVG justifiant d'un traitement par <medicament /> et la réduction de la posologie de <medicament /> à 1/jour*

6 Discussion et conclusion

La qualité de l'anonymisation à la source se révèle excellente en précision et son rappel est bon (sous-anonymisation). Nous sommes partis de cette version anonymisée pour constituer la référence ; à ce titre, il ne nous a guère été possible de mettre en évidence des sur-anonymisations. En revanche, nous avons pu pointer les éléments n'ayant pas été anonymisés. En pratique, un examen des balises nom et prénom montre une anonymisation totale des noms de famille et quasi-totale des prénoms. Cela confirme dans notre contexte la pertinence de la démarche d'anonymisation initiale qui assure déjà un bon niveau d'anonymisation. De plus, les prénoms ratés au premier niveau comportent des lettres accentuées, signe d'un problème sans doute simple à résoudre.

Deux éléments d'appréciation doivent être pris en compte dans le commentaire des résultats de notre portage partiel de DE-ID au français. En premier lieu, nous n'avons pas

été en mesure de produire autant de ressources linguistiques que celles existant en anglais (nous n'avons pas distingué les prénoms masculins et féminins ni même effectué une distinction entre prénoms ambigus et non ambigus ; il en est de même pour les noms de famille, d'hôpitaux, de lieux, etc). Pour les ressources qui nous faisaient défaut en français, nous avons repris celles existant en anglais s'il s'agit de ressources potentiellement existantes en français (par exemple des listes de noms et prénoms complémentaires). Nous avons abandonné les ressources anglaises le cas échéant. En second lieu, nous n'avons pas été en mesure de modifier efficacement les expressions régulières utilisées en anglais, ce qui se révèle source de nombreuses suranonymisations. Si les premiers points pourraient être améliorés à l'aide de ressources complémentaires, ce dernier point est plus bloquant : il demande de revoir le cœur du repérage d'expressions du programme. Cela justifie à notre avis le choix de reconstruire un nouvel anonymiseur, choix globalement plus efficace.

Pour des raisons de temps, seul un petit échantillon de notre corpus a été anonymisé manuellement. Un échantillon plus grand est en cours de constitution. En tout état de cause, dans la mesure où notre échantillon a été tiré aléatoirement et où les résultats obtenus sont tranchés, nous nous attendons à ce que l'échantillon plus large confirme nos premières mesures.

Avec un investissement jusqu'ici modéré dans son développement, Medina obtient des résultats honorables comme anonymiseur de second niveau. Ici aussi, l'emploi de ressources complémentaires devrait permettre d'augmenter le rappel. Par ailleurs, l'utilisation de composants de traitement automatique des langues (étiqueteur morphosyntaxique, boîte à outils de gestion de transducteurs à états finis, analyseur syntaxique, etc.) pourrait aider à améliorer la précision. Notons cependant que l'usage de tels outils demande généralement une adaptation[16].

Références

[1] Pestian JP, Brew C, Matykiewicz P, et al. A shared task involving multi-label classification of clinical free text. In : *Proceedings of BioNLP 2007*. Association for Computational Linguistics, 2007.

[2] Kim JD, Ohta T, Tateisi Y, and Tsujii J. Genia corpus-a semantically annotated corpus for bio-textmining. *Bioinformatics* 2003;19(Suppl. 1):180–2.

[3] Zweigenbaum P. Natural Language Processing in the medical and biomedical domains : a parallel perspective. In: Rebholz-Schuhmann D, Salakoski T, and Pyysalo S, eds, *Proceedings 3rd International Symposium for Semantic Mining in Biomedicine* (SMBM 2008), Turku. 2008; pp. 3–4. Keynote speech.

[4] Neamatullah I, Douglass MM, Lehman LWH, et al. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making* July 2008;8.

[5] Ruch P, Baud R, Rassinoux A, Bouillon P, and Robert G. Medical document anonymization with a semantic lexicon. *Proc AMIA Symp* 2000:729–33.

[6] Grouin C. *Chaîne de traitement pour la constitution automatique de corpus : application sur le domaine médical pour le projet corpus CLEF*. DESS d'ingénierie multilingue, Institut National des Langues et Civilisations Orientales,

[16] Le test d'un système de reconnaissance d'entités nommées « général » (utilisé dans un système de recherche de réponses à des questions en domaine ouvert) non modifié (hormis la conversion de ses balises de sortie) a donné des résultats très faibles, avec un rappel de 0,15.

2002.

[7] Zweigenbaum P, Jacquemart P, Grabar N, and Habert B. Building a text corpus for representing the variety of medical language. In : Patel VL, Rogers R, and Haux R, eds, *Medinfo*, 2001; pp. 290–4.

[8] El Emam K et Kamal-Dankar F. Protecting privacy using k-anonymity. *J Am Med Inform Assoc* September/October 2008;15(5):627–37.

[9] Uzuner O, Luo Y, and Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc* 2007;14:550–63.

[10] Friedlin JF, and McDonald CJ. A software tool for removing patient identifying information from clinical documents. *J Am Med Inform Assoc* September/October 2008;15(5):601–10.

[11] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *J Am Med Inform Assoc* 2001;8(suppl):17–21.

Adresse de correspondance

Cyril Grouin, LIMSI-CNRS, BP133, F-91403 Orsay Cedex, France

grouin@limsi.fr, http://www.limsi.fr/