# ETeRNAL 2015

# Table des matières

# Recherche des indices permettant une identification: l'anonymisation des transcriptions du corpus ESLO

Iris Eshkol-Taravella[1] Olivier Baude[1] Denis Maurel[2] Layal Kanaan-Caillol[1]

(1) LLL, UMR 7270, CNRS, UFR LLSH, 10 Rue de Tours 45065 ORLEANS cedex 2
(2) Université François-Rabelais de Tours, LI

iris.eshkol@univ-orleans.fr, olivier.baude@ univ-orleans.fr, denis.maurel@univ-tours.fr, layal.kanaan@univ-orleans.fr

**Résumé.** Cet article aborde la question de l'anonymisation automatique des corpus oraux afin de permettre leur utilisation et diffusion sur la Toile. Nous proposons une analyse des éléments constituant un « faisceau d'indices » qui, dans un certain contexte, contribue à l'identification. Ces indices dépassent par leur diversité et leur hétérogénéité les entités nommées. Nous décrivons ensuite une expérimentation du repérage automatique de ce faisceau d'indices dans les transcriptions.

## Abstract.

**Recognizing clues leading to identification: anonymizing the transcriptions of the ESLO speech corpus**
This article tackles the question of oral corpus anonymization in preparation for its diffusion on the Web. We first analyze elements constituting a « clues set » which contribute to the identification. Those clues exceed named entities by their diversity and heterogeneity. Then we describe an experiment based on a module of automatic recognition of its clues in the transcriptions.

**Mots-clés :** anonymisation, anonymisation automatique, corpus oral, faisceau d'indices, données personnelles, identification
**Keywords:** anonymisation, automatic anonymisation, oral corpus, indications set, personal data, identification

## 1 Introduction

Grâce au développement des outils informatiques, la mise à disposition de différents corpus a modifié le travail des chercheurs en linguistique, en sciences sociales et humaines et en traitement automatique des langues (TAL). Les initiatives actuelles se développent autour de la diffusion et de la disponibilité de ces ressources en accès - souvent libre - sur la Toile. Les corpus oraux en langues étrangères le BNC[1], le Russian National Corpus[2] ou encore le National Corpus of Polish[3], ou en français, CLAPI[4], PFC[5], CRFP[6], Corpus de la parole, etc. sont apparus sur le Toile et plus récemment la France s'est doté d'un EQUIPEX dédié à la diffusion des ressources linguistiques (EQUIPEX ORTOLANG). Pour diffuser ces corpus, les questions juridiques dont celle de leur anonymisation se sont avérées primordiales.

[3] British National Corpus, http://www.natcorp.ox.ac.uk/
[2] http://www.ruscorpora.ru/en/index.html
[3] http://nkjp.pl/index.php?page=0&lang=1
[4] Corpus de langues parlées en interaction, http://clapi.univ-lyon2.fr/
[5] Phonologie du français contemporain, http://www.projet-pfc.net/?accueil:intro
[6] Corpus de référence du français parlé, http://www.up.univ-mrs.fr/delic/crfp

La linguistique sur corpus oraux a bénéficié d'un travail précurseur pour la collecte et la diffusion d'enregistrements sonores et de leurs transcriptions. Sous l'égide du Ministère de la Culture et du CNRS un groupe de travail constitué de linguistes, d'informaticiens, de juristes et de conservateurs a réfléchi aux aspects juridiques et éthiques de l'usage des corpus oraux. Ce travail s'est concrétisé par la publication de l'ouvrage *Corpus oraux, guide des bonnes pratiques 2006* (Baude et al., 2006). L'anonymisation est une pratique qui répond à un impératif juridique précis. Sans recueil du consentement de la personne enregistrée, il est obligatoire d'empêcher son identification. L'impossibilité d'identifier est une notion complexe qu'on a trop souvent réduite à l'effacement des noms propres. La tâche est bien plus difficile, mais aussi plus stimulante pour les recherches en linguistique et en TAL.

L'anonymisation relève de procédures différentes selon qu'on traite l'enregistrement sonore, sa transcription ou les métadonnées descriptives. Toutefois, dans tous les cas, l'objectif reste le même. Si selon certains juristes la voix est une donnée identifiante ce qui nécessiterait de modifier le signal acoustique de tout enregistrement et par là même obèrerait toute recherche en linguistique, les pratiques des chercheurs s'orientent plus généralement vers un traitement des données personnelles au sens large. Que ce soit sur l'oral ou sur l'écrit celles-ci sont diverses, il peut s'agir d'une forme nominative, d'une profession, d'un statut, d'une caractéristique physique, etc. et/ou du recoupement de plusieurs de ces informations. Si l'on convient que l'anonymisation ne se réduit pas à l'effacement des noms propres, il est nécessaire de définir avec précision quels sont les traitements à effectuer pour répondre à l'objectif de réduire les possibilités d'identification. Dans le cas de grands corpus, ces traitements deviennent une étape fondamentale du travail de constitution du corpus avec des effets très importants sur la gestion et la diffusion des données.

Le travail décrit dans cet article porte sur le corpus oral ESLO (Enquête Sociolinguistique à Orléans). Il s'agit d'un grand corpus de données orales qui regroupe deux enquêtes ESLO 1 et ESLO 2 (Baude, Dugua, 2011, Eshkol-Taravella et al., 2012). ESLO 1 a été réalisé entre 1968 et 1974, à l'initiative d'universitaires britanniques avec une visée didactique : l'enseignement du français langue étrangère dans le système public d'éducation anglais. Il a été numérisé et transcrit par l'équipe du Laboratoire Ligérien de Linguistique (LLL). ESLO2 est une nouvelle enquête, débutée en 2008 par le LLL. Réunis, ESLO 1 et ESLO 2 forment une collection de 700 heures d'enregistrement (10 millions de mots), ce qui est considéré aujourd'hui comme une valeur repère pour les investigations projetées. Il s'agit en somme d'un très grand corpus dont l'objectif de mise à disposition a déclenché une réflexion sur les éléments permettant l'identification du locuteur et de toute autre personne mentionné dans le discours de celui-ci et sur leur repérage automatique.

# 2 Identification à travers un « faisceau d'indices »

Selon le Dictionnaire d'analyse du discours, « l'identité résulte, à la fois, des conditions de production qui contraignent le sujet, conditions qui sont inscrites dans la situation de communication et/ou dans le préconstruit discursif, et des stratégies que celui-ci met en œuvre de façon plus ou moins consciente » (Charaudeau, Maingueneau, 2002:300). Les auteurs distinguent une identité psychosociale consistant en traits qui définissent le sujet selon son âge, son sexe, son statut, etc. et une identité discursive du sujet énonciateur « qui peut être décrite à l'aide de catégories locutives, de modes de prise de parole, de rôles énonciatifs et de modes d'interventions » (ib.) Nous n'allons pas nous intéresser, dans cette étude, aux stratégies discursives que choisit le sujet parlant pour se construire une identité : sa manière de prendre la parole, de thématiser ses propos, d'organiser son argumentation. Notre objectif est d'étudier, dans le discours oral, des éléments qui permettent de distinguer le sujet parlant et la personne dont on parle des autres et, par conséquent, de les reconnaître. Nous avons appelé l'ensemble de ces éléments un « faisceau d'indices ». On peut identifier la personne en la dénommant, c'est-à-dire en la mentionnant par son nom, ou en la décrivant, c'est-à-dire en représentant certains de ses traits, voire de ses activités. Anonymiser le corpus consiste dans le repérage de ces indices et leur substitution par un hyperonyme ou un élément à référents multiples. Ces indices peuvent être de nature lexicale et sémantique très variée : des entités nommées (section 2.1), d'une part, mais aussi des groupes nominaux fondés sur le nom commun désignant les traits caractéristiques ou des groupes verbaux, énoncés décrivant les habitudes et les activités sociales de la personne (section 2.2).

## 2.1 Entités nommées identifiantes

Traditionnellement la tâche d'anonymisation s'arrête au repérage des entités nommées (noms de personnes, lieux, organisations, âges, etc.) (Ehrmann, 2008, Nadeau, Sekine, 2004). C'est le cas de plusieurs travaux en TAL dans le domaine médical (Meyster et al., 2010, Tweit et al., 2004, Raaj, 2012, Uzuner et al., 2007, Grouin, Zweigenbaum, 2011) qui portent sur les documents écrits (rapports, dossiers médicaux, etc.) et où les informations à anonymiser sont assez homogènes et souvent regroupées dans un endroit précis. Un de ses outils disponible gratuitement est Medina

(Medical Information Anonymization[7]). Il repère automatiquement à l'aide de patrons et de lexiques les noms de personnes, les lieux, les noms d'hôpitaux et les informations numériques comme les adresses, âges, numéros de téléphones, etc. dans les documents cliniques en français.

Les entités nommées sont effectivement les candidates idéales à l'anonymisation car par définition, « on appelle entité nommée tout expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus » (Ehrmann, 2011), les entités nommées sont donc censées renvoyer vers un référent unique. Or il s'avère que tous les entités nommées mentionnées dans le discours ne sont pas identifiantes du point de vue de l'anonymisation.

**- Noms de personnes**

Le premier cas des entités nommées est les noms de personne. C'est un indice fort pour l'anonymisation car le rôle même de ces noms est de nommer, c'est-à-dire d'indiquer le référent d'une personne mentionnée. Cela est prouvé par l'anonymisation manuelle des transcriptions (section 3). Dans un test sur 112 fichiers d'ESLO1, 168 éléments ont été masqués. Parmi eux, les 159 éléments ont été remplacés par hyperonyme *NPERS.*

Les noms de personne jouent aussi le rôle primordial dans les travaux sur la détection de l'identité du locuteur dans les journaux télévisés (Charhad, Quénot, 2005). Les auteurs reconnaissent le locuteur grâce aux patrons qui détectent la personne qui se présente, qui vient de parler (le locuteur remercie, par exemple, l'orateur précédent en l'appelant par son nom) et la personne qui va parler (le locuteur passe la parole à un autre orateur en le nommant).

Est-ce que tous les noms évoqués dans le discours pointent vers l'identité de la personne ? Les noms de famille ou prénoms rares comme *Eshkol* ou *Kanaan,* dans le cadre de la ville comme Orléans, peuvent éventuellement identifier la personne. Pourtant, dans le cadre de l'anonymisation, les noms de personnalités *Sarkozy, Cotillard*, étant les noms publics ne doivent pas être anonymisés. Les noms de famille très répandus qui renvoient à un nombre élevé de référents comme *Dupont, Durand* ne donnent aucune information sur la personne et ne permettent pas à eux-seuls de l'identifier.

**- Fonction**

Selon le guide d'annotation des entités nommés Quaero (Rosset et al. 2011), la fonction (*func*) comprend les métiers, les fonctions et les rôles sociaux de la personne.

Nommer la personne par sa fonction *maire d'Orléans, directeur du collège de Saint-Jean de Braye* est un acte qui peut renvoyer à un référent unique. On est de nouveau en présence d'un indice fort. Ce n'est pourtant pas le cas d'un nom de métier. La mention dans le discours du métier *enseignant-chercheur* ne veut rien dire sur l'identité de la personne, mais dans le contexte *je suis enseignant-chercheur* il devient un indice de l'identification.

**- Autres entités nommées**

Les autres entités nommées présentes dans le discours doivent aussi avoir un lien avec le locuteur ou la personne qu'il mentionne dans son discours pour devenir un indice d'identification. Ce lien est souvent exprimé dans le discours même par le contexte gauche/droite de l'entité ou par la question posée dans le cadre de l'entretien ce qui est le cas du corpus étudié. Ainsi, le nom de lieu tout seul ne dit rien sur la personne mais employé avec la précision *je travaille à ….* ou *mon père est originaire de…,* il devient un indice, une information personnelle. Il le devient aussi dans les réponses à des questions portant sur l'identité de la personne comme *où travaille votre femme ? vous êtes originaire d'où ?*. Les mêmes observations peuvent se faire pour d'autres types d'entités nommées : les dates ou encore les noms d'organisations.

De manière concomitante, il y a dans le discours d'autres éléments qui ne font pas partie des entités nommées mais qui peuvent renvoyer vers l'identité de la personne. Ce phénomène a été déjà mentionné dans (Amblard, Fort, 2014) où les auteurs présentent entre autres le processus d'anonymisation automatique du discours transcrit de schizophrènes. Ils notent l'insuffisance du simple repérage à l'aide de scripts Python des mots commençant par une majuscule dans les extrais du corpus où des sujets relatent un événement « s'inscrivant dans une temporalité et une géographie particulière » et la présence d'autres indices selon lesquelles on peut identifier le locuteur ou ses proches. Cette affirmation se manifeste à travers les chiffres provenant des résultats de l'expérience de l'annotation automatique du sous-corpus ESLO1 en indices permettant l'identification éventuelle du locuteur (section 4). Dans 112 fichiers de transcription d'ESLO1 annotés en entités nommées et en indices, candidats à l'anonymisation, on retrouve 13 909 entités nommées au total et seulement 1 038 autres indices. Ces chiffres confirment que, d'une part, toutes les entités nommées ne

[7] http://medina.limsi.fr/

renvoient pas vers le locuteur et que, d'autre part, il existe dans le corpus d'autres éléments qui peuvent permettre l'identification éventuelle de la personne.

### 2.2 Autres indices

Si l'on veut anonymiser efficacement le discours, on ne peut pas s'arrêter aux entités nommées car d'autres indices peuvent renvoyer vers le locuteur ou vers la personne dont il parle. Observons les exemples tirés du corpus ESLO1 :

- *j'ai une maladie du foie ça m'a même occasionné une petite scoliose déformation légère de la colonne vertébrale.*
- *mon père a fondé un le plus grand cabinet d'ophtalmologiste de la ville*[8]
- *je suis scout de France le jeudi soir où j'anime un un atelier photos*[9]

Cette catégorie des indices est large. Elle inclut des éléments assez hétérogènes désignant les différentes informations personnelles sur la personne : événements, activités sociales, loisirs, maladies, handicap, etc. qui peuvent au même titre que le travail, la famille donner les informations sur le locuteur ou la personne dont on parle.

Ainsi, le « faisceau d'indices » inclut les entités nommées identifiantes, mais peut contenir aussi d'autres éléments qui permettent l'identification soit directement, soit, par combinaison au sein de ce faisceau : la personne est patron d'un bar au moment d'enregistrement, et avant elle travaillait dans l'aviation militaire. Le processus d'identification est progressif, il se construit au fur et à mesure de l'accroissement des indices. On peut supposer qu'un indice identifiant ou une série de ces indices est associée à un individu particulier dans la mémoire à l'aide d'un certain lien dénominatif qui sera réactivé lors de leur apparition dans le discours. C'est grâce aux facteurs contextuels, c'est-à-dire grâce aux connaissances que l'utilisateur du corpus maîtrise concernant le locuteur ou la personne mentionnée dans le discours de celui-ci, que l'identification peut se faire.

Les parties qui suivent sont consacrées à la description de l'anonymisation du corpus ESLO. Le processus d'anonymisation du corpus consiste à repérer un faisceau d'indices qui permet d'identifier le sujet parlant ou toute autre personne mentionnée dans le discours. Dans le processus actuel de l'anonymisation des transcriptions d'ESLO (section 3), ces indices sont repérés manuellement par les transcripteurs. Pour aider ce processus, une expérimentation de l'automatisation de ce processus a été tentée (section 4). Nous finirons par quelques perspectives liées à l'intégration du module automatique développé dans le processus actuel (section 5).

## 3 Procédure semi-automatique d'anonymisation des transcriptions dans le corpus des ESLO

Nous allons voir dans cette partie la procédure suivie par les gestionnaires du corpus des ESLO afin de procéder à l'anonymisation des données du corpus.

Du point de vue juridique, le corpus ESLO1, a posé deux problèmes (Baude[10]). Premièrement, les locuteurs n'ont rempli aucun document pour exprimer leur consentement ; deuxièmement, les locuteurs de la fin des années soixante ne pouvaient pas prévoir que leurs enregistrements pourraient être diffusés par Internet qui n'existaient pas à l'époque. Dans le cas d'ESLO2, les locuteurs signent un document de consentement à la diffusion de l'ensemble des données brutes. Le choix de l'équipe a néanmoins été d'anonymiser l'ensemble des données d'ESLO1 et d'ESLO2.

L'anonymisation actuelle dans ESLO est semi-automatique et porte sur deux types d'objets : les données (sons et transcriptions) et les métadonnées. Dans la chaîne de traitement du corpus, la phase d'anonymisation est fractionnée ;

[8] L'emploi des déterminants *un le* dans cet énoncé fait partie des disfluences de l'oral (autocorrection) et est transcrite comme telle dans les fichiers de transcription.

[9] idem.

[10] http://eslo.huma-num.fr/index.php/pagemethodologie?id=69

elle précède la phase de transcription, coïncide avec elle et lui succède. Nous nous contentons, dans cet article, de décrire la phase de d'anonymisation des transcriptions[11].

Le codage des noms propres des locuteurs est l'action la plus classique et attendue dans une procédure d'anonymisation. Les transcriptions comportent des informations issues des métadonnées, à savoir l'identifiant du locuteur. Dans la procédure ESLO, des codes aléatoires sont générés par l'application suite à la création d'une fiche en saisissant les métadonnées du locuteur (ex : DC738). Ces codes sont repris dans les transcriptions. Le traitement des données identifiantes contenues dans les énoncés est effectué au niveau-même de la transcription. Il est demandé aux transcripteurs de remplacer par l'hyperonyme *NPERS* les noms de personnes (Figure 1) et par *NANON*[12] les autres segments du discours permettant d'identifier un locuteur.

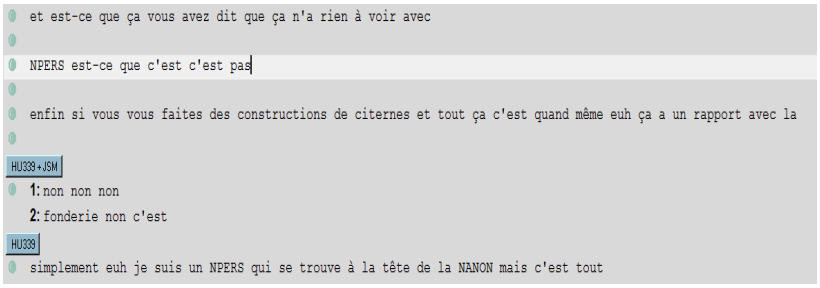


Figure 1 : Anonymisation dans la transcription

L'anonymisation manuelle des fichiers de transcription a soulevé la question d'automatisation de ce processus grâce aux outils du TAL. L'expérience a été menée afin de repérer automatiquement les indices permettant l'identification éventuelle du locuteur ou de toute autre personne mentionnée dans les transcriptions.

# 4 Expérience de l'anonymisation automatique sur un sous-corpus d'ESLO1

L'expérimentation décrite dans cette partie a été effectuée en collaboration avec le laboratoire LI (Laboratoire Informatique) de l'université de Tours. Le test portait sur un sous-corpus d'ESLO1 (112 entretiens face-à-face) contenant de nombreuses données personnelles sur le locuteur car il s'agissait d'un questionnaire concernant la vie des témoins : « *Depuis combien de temps habitez-vous Orléans ? » « Quel âge avez-vous ? » « Qu'est-ce que vous faites comme métier ? » « Où travaillez-vous ? » « Qu'est-ce que fait votre époux(se) ? »,* etc.

## 4.1 Repérage automatique

Lorsqu'on parle de l'anonymisation automatique tout le monde s'accorde sur la nécessité de repérer les entités nommées. Comme nous l'avons évoqué, ces éléments font très souvent partie, à juste titre, des indices recherchés. C'est la raison pour laquelle, pour faire une expérimentation d'anonymisation automatique des transcriptions d'ESLO1, il a été décidé de partir de l'outil permettant d'identifier les entités nommées. La collaboration avec le LI de Tours a permis d'exploiter le système CasSys développé dans le cadre de la thèse par Nathalie Friburger (Friburger, 2002) et intégré à la plate-forme Unitex (Paumier, 2003). Il s'agit d'une approche symbolique en surface permettant de construire les grammaires locales selon le contexte sous forme des cascades de transducteurs qui repèrent et annotent les entités nommées dans le discours médiatique.

Le système CasSys a été adapté au corpus traité. Tout d'abord, le corpus a été segmenté en tours de parole en fonction des balises Transcriber[13]. Les cascades de CasSys ont été ensuite enrichies de nouvelles grammaires locales avec des dictionnaires et des graphes spécifiques pour reconnaître dans les transcriptions de l'oral en plus des entités nommées

[11] Pour une présentation de la procédure : Baude et Dugua Guide d'Anonymisation (en ligne http://eslo.huma-num.fr/index.php/pagemethodologie?id=69)

[12] Nom anonymisé

[13] Méthode recommandée par (Dister, 2007)

d'autres indices. Enfin, en tenant compte de la nature du corpus, les différentes disfluences de l'oral ont été prises aussi en compte comme par exemple dans *je m'appelle euh Patrick Mallon*[14].

Nous avons procédé en deux étapes. Tout d'abord, nous lançons des cascades de transducteurs qui repèrent et annotent les entités nommées (*EN*). Ensuite, une autre série de cascades appliquée à ce corpus annoté, identifie les indices-candidats à l'anonymisation (*DE*[15]). Dans cet exemple, l'entité nommée *Pithiviers* a été reconnue au cours de la première étape, cette entité devient identifiante au cours de la deuxième étape car elle se trouve dans le contexte indiquant son lien avec le locuteur *moi je suis native de Pithiviers* :

- 1ère étape : <***EN*** *type="loc.admi">**Pithiviers**</**EN**>*
- 2ème étape : <***DE*** *type="pers.speaker">**moi je suis** <**DE** type="identity.origin">**native de** <EN type="loc.admi">Pithiviers</EN></**DE**></**DE**>*

La Figure 1 présente le graphe permettant la reconnaissance d'une origine géographique : ce graphe appelle un sous-graphe (*NELoc*) qui reconnaît un toponyme identifié par la cascade des entités nommées.

Suite à l'analyse manuelle du corpus et à partir de la typologie de la campagne d'évaluation Ester2 (campagne d'évaluation des systèmes de transcription enrichie d'émissions radiophoniques)[16] nous avons élaboré le jeu d'étiquettes pour annoter des indices. L'enquête correspond essentiellement à des questions concernant la personne interrogée et sa famille : origine, âge, naissance, arrivée à Orléans, travail et même syndicat. Pour cela nous avons défini une typologie avec trois types principaux, personne, identité et travail, eux même divisés en sous-types, comme présenté dans la Figure *3*. Le sujet sur qui porte l'information est annoté en premier lieu. Nous distinguons entre le locuteur (*pers.speaker*) et les autres membres de sa famille (*pers.spouse*, *pers.parent*, *pers.child*). Nous précisons ensuite la nature de cette information : l'identité, le travail, les études, l'engagement associatif ou syndicale, les vacances :

- *il est parti à Paris =>*
  *<**DE** type="pers.child">**il est parti** <**DE** type="work.location">**à** <**ENT** type="loc.admi">**Paris**</**ENT**></**DE**> il travaille dans les <Sync time="1526.195"/> <**DE** type="work.field">dans les assurances</**DE**></**DE**>*
- *alors je suis monsieur Gabrion je suis ingénieur chimiste=>*
  *alors **<DE** type="pers.speaker"><**DE** type="identity.name">**je suis** <**ENT** type="pers.hum">**monsieur Gabrion**</**ENT**></**DE**></**DE**>* *<**DE** type="pers.speaker">**je suis** <**DE** type="work.occupation">**ingénieur chimiste**</**DE**></**DE**>*
- *je peux vous demander quel est votre syndicat ? </Turn> <Turn speaker="spk5" startTime="5071.106" endTime="5072.466"> <Sync time="5071.106"/> <Sync time="5071.22"/> oui c'est la <**DE** type="pers.speaker"> <**DE** type="involvement.tradeunion"> <**ENT** type="org"> **CGT** </**ENT**></**DE**></**DE**>*
- *de ce fait <**DE** type="pers.speaker">**nous sommes allés euh** <**ENT** type="time.date.rel"> **trois jours** </**ENT**> <**DE** type="trip.work"> **à** <**ENT** type="loc.admi"> **Londres** </**ENT**> </**DE**> <**ENT** type="time.date.rel"> **trois jours** </**ENT**> <**DE** type="trip.work"> **à** <**ENT** type="loc.admi"> **Vienne** </**ENT**> </**DE**> **nous avons été** <**ENT** type="time.date.rel"> **trois jours** </**ENT**> <**DE** type="trip.work"> **en** <**ENT** type="loc.admi">*
  ***Hongrie** </**ENT**></DE>*

On voit dans ces exemples, que l'entité nommée *monsieur Gabrion* est bien annotée en tant qu'indice car elle se trouve dans un contexte qui concerne le locuteur *je suis monsieur Gabrion*. C'est le cas pour une autre entité nommé, le nom du syndicat *CGT*, car elle se trouve dans la réponse à la question concernant le locuteur. Les cascades annotent également les syntagmes fondés sur les noms communs comme le métier *ingénieur chimiste* ou le domaine d'activités professionnelles *dans les assurances*. A cela s'ajoute l'annotation d'autres indices comme les vacances *nous sommes allés trois jours à Londres* ou des autres actions *il est parti de Paris*.

La reconnaissance des indices est fondée sur le contexte qui joue un rôle primordial dans le processus d'identification car il permet de réduire le champ d'application de ces éléments à un seul individu, de le distinguer des autres référents possibles. En premier lieu, on peut mentionner le contexte immédiat (gauche et/ou droite) d'indice. Le nom de lieu

[14] L'annotation automatique des entités nommées et dénommantes a été décrite dans (Maurel et al., 2011, Eshkol et al., 2012).

[15] Le terme que nous avons utilisé à l'époque de cette expérimentation pour désigner les indices annotés est celui d'« entité dénommante » (Eshkol, 2010). Le travail complémentaire entrepris depuis sur la définition de cette notion nous amène maintenant à préférer la terminologie de « faisceau d'indices » telle que nous l'avons présentée dans la section 2.

[16] http://www.afcp-parole.org/ester/docs/Conventions_EN_ESTER2_v01.pdf

n'aura pas grand intérêt employé seul, mais employé avec des verbes comme *venir de*, *travailler à* ou avec des noms comme *collège, hôpital*, *etc.* il devient identifiant du lieu de travail, d'études ou d'origine de la personne. L'indice repéré doit être étiqueté aussi selon le contexte. Dans la phrase *je travaille au collège de Saint-Jean-de-Braye,* l'entité nommée *collège de Saint-Jean-de-Braye* ne réfère plus seulement à un établissement scolaire en général, c'est une référence à un lieu de travail du locuteur. Ce contexte peut être aussi défini par la question posée. On sort ce faisant des limites de l'énoncé pour étudier un contexte plus large. Le nom de lieu, par exemple, n'est pas signifiant s'il est utilisé pour répondre à la question : *où parle-t-on le mieux le français ?*, par contre il devient un indice dans les réponses aux questions concernant les origines du locuteur, ou dans les énoncés décrivant l'emploi du locuteur, pour autant que celui-ci indique le lieu de son travail. De la même manière, les réponses aux questions sur les émissions de télévision, par exemple, n'apportent pas d'information personnelle et les noms de personnes qui apparaissent n'ont pas à être pris en compte. Les questions posées peuvent donc jouer un rôle important dans la catégorisation adéquate d'un indice repéré. Enfin, il est nécessaire de prendre en compte le contexte socioculturel de l'époque. Ainsi, les destinations de vacances peuvent être prises en compte car en 1968 peu de gens à Orléans voyageaient à l'étranger, c'est le cas du dernier exemple ci-dessus.

L'annotation a été réalisée sur 112 fichiers Transcriber (35,75 Mo). L'évaluation des résultats a été effectuée sur 9 fichiers (6 fichiers ont été réservés pour les tests). Les indices ont été reconnus avec la précision estimée à 94,2 % et le rappel de 84,4 % (Maurel et al., 2009).

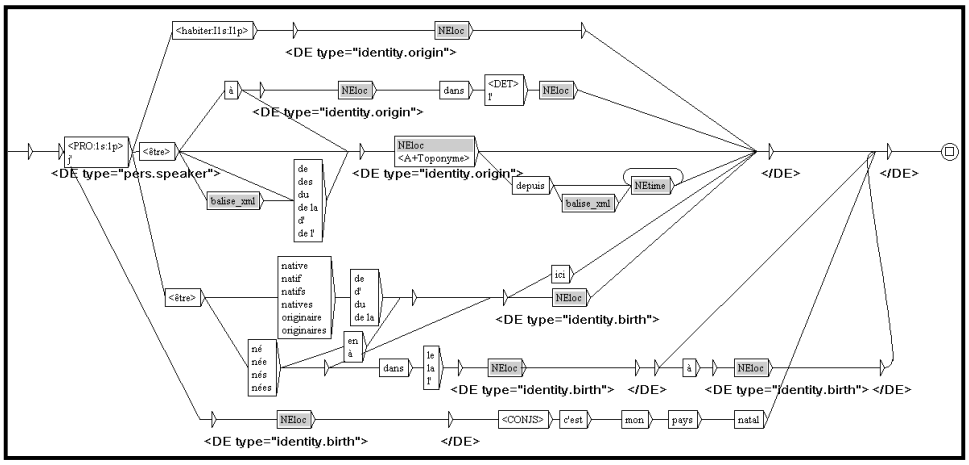


Figure 2 : Un graphe pour l'origine géographique

## 4.2 Difficultés rencontrées

Malgré ce succès, plusieurs difficultés ont été mises en évidence.

En premier lieu, la présence de multiples disfluences (hésitations, répétitions, reformulations, amorces, etc.) qui peuvent intervenir à différents moments dans le discours comme dans *je m'appelle euh Patrick Mallon* rendent la tâche difficile. Le graphe contenant la liste de disfluents possibles a été créé ce qui a permis de résoudre ce problème dans beaucoup de cas.

Ensuite, dans le discours oral, les informations apparaissent d'une manière parfois aléatoire. Ainsi, des informations sur le témoin ne se trouvent pas nécessairement dans la partie questionnaire, mais peuvent surgir à des endroits inattendus. Par exemple, la description de la recette de l'omelette peut être l'occasion de glisser son origine géographique :

– *enfin on assaisonne sel poivre euh <DE type="pers.speaker"> nous en <DE type="identity.origin"> <ENT type="loc.admi"> Lorraine </ENT></DE></DE> on on découpe des petits des petits morceaux de lards qu'on fait frire avant*

Certaines informations doivent être aussi parfois déduites du contexte comme dans l'exemple suivant:

*BV: y a longtemps que vous êtes à Orléans ?*
*MS530:euh oui euh vingt-deux ans*
*BV: ça fait euh vous êtes née à Orléans*
*MS530: oui*

Une autre difficulté provient de la variation linguistique. Les informations de nature personnelle varient d'une manière non homogène dans le corpus. Chaque type d'information peut être présenté à travers un groupe nominal ainsi qu'avec des expressions plus étendues. Ainsi, le locuteur peut décrire son métier de manières diverses :

– *je suis enseignant dans l'école publique*
– *je suis maître auxiliaire*
– *j'enseigne des mathématiques modernes des mathématiques classiques de la chimie et de la technologie*

| | |
|---|---|
| Personne (+pers) | la personne interrogée (+speaker) |
| | son conjoint (+spouse) |
| | ses enfants (+child) |
| | les autres membres de la famille (+parent) |
| Identité (+identity) | le nom (+name) |
| | l'adresse (+addr) |
| | l'âge (+age) |
| | le mariage (+wedding) |
| | l'origine (+origin) |
| | la naissance (+birth) |
| | l'arrivée à Orléans (+arrival) |
| | le nombre d'enfants (+children) |
| Travail (+work) | métiers (+occupation) |
| | secteur d'activité (+field) |
| | lieu de travail (+location) |
| | entreprise (+business) |
| Engagement (+involvement) | association (+voluntary) |
| | militaire (+military) |
| | scolaire (+school) |
| | syndical (+tradeunion) |
| Voyage (+trip) | études (+study) |
| | vacances (+holiday) |
| | professionnel (+work) |
| Etudes (+study) | lieu (+location) |
| | diplôme (+degree) |
| | établissement (+edu) |

Figure 3 : Typologie des indices

On ne peut jamais atteindre une liste exhaustive de toutes les reformulations possibles.

Enfin, le corpus peut comprendre des informations difficiles à catégoriser comme par exemple :

- *mon père a fondé un le plus grand cabinet d'ophtalmologiste de la ville*
- *je suis scout de France le jeudi soir où j'anime un un atelier photos*

Cette catégorie du faisceau d'indices comprenant les actions, les événements, les activités sociales du locuteur semble « imprévisibles » en raison de son manque d'homogénéité.

Malgré toutes ces difficultés, les indices peuvent être reconnus automatiquement avec une bonne précision et un bon rappel. La première catégorie, les entités nommées identifiantes, est bien reconnue par le module développé. La deuxième indiquant les actions, événements, activités sociales du locuteur est identifiée mais pas d'une manière exhaustive.

Cependant la multitude d'éléments personnels annotée dans le corpus soulève une autre question concernant leur pertinence. Tous les éléments annotés ne nécessitent pas d'être anonymisés. Actuellement, la décision d'anonymiser un tel ou tel indice, ne peut se faire que manuellement par un humain. C'est seulement l'humain qui peut décider aujourd'hui, souvent d'une manière assez subjective, lequel des éléments personnels renvoie le plus vers le locuteur ou ses proches et doit donc être masqué. Le principe respecté est de garder le maximum d'informations pour pouvoir permettre l'analyse du corpus. Ainsi, dans les 112 fichiers contenant 1 038 indices annotés, seulement 168 ont été remplacés par leur hyperonyme (159 *NPERS* et 9 *NANOM*[17]).

# 5 Conclusion et perspectives

Le travail effectué a montré que si l'on veut anonymiser un corpus d'enquêtes sociolinguistiques, il ne suffit pas de reconnaître les noms propres et les autres entités nommées car d'une part, d'autres éléments peuvent aussi permettre l'identification du locuteur ou de la personne mentionnée dans le discours notamment quand il existe une combinaison de ces éléments au sein du corpus et, d'autre part, tous les entités nommées ne sont pas sensibles à l'anonymisation et ont besoin d'un contexte pour devenir identifiantes.

Le module développé pour le repérage automatique des indices-candidats à l'identification potentielle de la personne tient compte des spécificités de l'oral (la présence de disfluences, l'absence des signes de ponctuation dans les transcriptions, la segmentation en tours de parole) et permet d'obtenir des résultats encourageants.

La difficulté majeure de l'anonymisation automatique des discours transcrits de l'oral est que toutes les informations personnelles n'identifient pas la personne mais qu'en revanche une combinaison de certaines d'entre elles constituent un faisceau qui dans un certain contexte, le plus souvent extralinguistique, contribuent à l'identification. Actuellement la décision sur la pertinence de masquer certains éléments du faisceau ne peut se faire que par une intervention humaine. Pour aider cette validation manuelle, la distinction pourrait se faire entre les éléments les plus sensibles à l'anonymisation, c'est-à-dire ceux qui apportent une information plus importante et plus spécifique, et ceux qui sont plus généraux. Ainsi, pour distinguer entre les noms de famille rares comme *Eshkol* ou *Kanaan* et très répandues *Dupond* ou *Durand,* on pourrait s'appuyer, dans le cas du corpus des ESLO, sur une information concernant la fréquence d'un nom propre, éventuellement pondérée par des critères géographiques. De la même manière, le locuteur peut désigner son métier par un seul mot *enseignant* ou en précisant *professeur de physique.* Ce passage d'un seul nom à un groupe nominal plus étendu grâce aux modifieurs « se manifeste par l'ajout de propriétés supplémentaires à la classe présentée par le groupe nominal minimal, ce qui diminue l'extension de la classe et rapproche le groupe d'une référence plus individualisante » (Eshkol, 2010 : 258). Ce processus concerne n'importe quelle caractéristique (maladie, loisir, etc.). On pourrait ainsi attribuer plus de poids à ces éléments sensibles à l'anonymisation ce qui diminuerait le nombre des indices candidats à l'anonymisation et de cette manière aiderait la validation manuelle.

Dans le faisceau d'indices, la deuxième catégorie comprenant les éléments ne faisant pas partie des entités nommées comme actions, événements, activités sociales, doit être approfondie d'autant plus qu'elle permet d'apporter une information sur le profil sociologique du locuteur. Le module développé tient compte de ces indices mais leur liste n'est pas exhaustive. Pour un travail futur, nous envisageons d'étudier avec précision dans le corpus ESLO, tous les éléments anonymisés par une procédure manuelle afin d'affiner la typologie du faisceau d'indices.

[17] L'étiquette *NANOM* signifie le nom anonymisé.

# Références

AMBLARD M., FORT K. (2014). Étude quantitative des disfluences dans le discours de schizophrènes : automatiser pour limiter les biais. Actes de *TALN2014*, Marseille, France.

BAUDE O. (2006). *Corpus oraux : guide des bonnes pratiques*. CNRS-Editions et Presses universitaires d'Orléans, 2006.

BAUDE O., DUGUA C. (2011). (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ?, vol. 10, *Corpus, Varia*.

CHARAUDEAU P., MAINGUENEAU D. (2002). *Dictionnaire d'analyse du discours*. Paris, Éditions du Seuil.

CHARHAD M., QUENOT G. (2005). Approche par patrons linguistiques pour la détection automatique du locuteur : application à l'indexation par le contenu des journaux télévisés. *Compression et Représentation des Signaux Audiovisuels (CORESA'05)*, Rennes.

DUBOIS J. (1973). *Dictionnaire de linguistique.* Paris, Larousse.

EHRMANN M. (2008). Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation. Thèse de doctorat, Université Paris 7 - Centre de recherche Xerox, Grenoble (XRCE).

ESHKOL I. (2010a). Entrer dans l'anonymat. Etude des « entités dénommantes » dans un corpus oral. *Eigennamen in der gesprochenen Sprache,* 245-266.

ESHKOL I., MAUREL D., FRIBURGER N. (2010b). Eslo: from transcription to speakers' personal information annotation. Actes de *Seventh Language Resources and Evaluation Conference (LREC 2010)*, Malte.

ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C., TELLIER I., (2012). Un grand corpus oral « disponible » : le corpus d'Orléans 1968-2012. *Ressources linguistiques libres*, *TAL*. 52 : 3, 17-46.

FRIBURGER N. (2002). *Reconnaissance automatique des noms propres ; application à la classification automatique de textes journalistiques*. Thèse de doctorat d'informatique, Université François Rabelais Tours.

GROUIN C, ZWEIGENBAUM P. (2011). Une approche à plusieurs étapes pour anonymiser des documents médicaux. *RSTI-RIA*, 25 :4, 525-549.

HAMON P. (1977). Pour un statut sémiologique du personnage. *Poétique du récit.* Barthes R. *et alii*, Points-Seuil, Paris.

MAUREL D., FRIBURGER N., ESHKOL I. (2009). Who are you, you who speak? Transducer cascades for information retrieval. Actes de *4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznań, Poland, 220-223.

MAUREL D., FRIBURGER N., ANTOINE J.-Y., ESHKOL-TARAVELLA I., NOUVEL D., (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *Varia TAL*, 52 :1, 69-96.

MEYSTRE S., FRIEDLIN B S., SHUYING S., SAMORE M. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology* 10.70.

NADEAU N., SEKINE S. (2009). *A survey of named entity recognition and classification*, Satoshi Sekine and Elisabete Ranchhod, ed., John Benjamins publishing company, 3-28.

PAUMIER S. (2003). *De la Reconnaissance de Formes Linguistiques à l'Analyse Syntaxique*. Thèse de Doctorat en Informatique, Université de Marne-la-Vallée.

RAAJ N. (2012). *Automated Tool for Anonymization of Patient Records.* Report. MSc Computing and Management, Imperiel College, London[18].

[18] http://www.comp.leeds.ac.uk/mscproj/reports/1112/raaj.pdf

ROSSET S., GROUIN C., ZWEIGENBAUM P. (2011). *Entitées nommées structurées : guide d'annotation Quaero.* Notes et documents LIMSI N°2011-04.

TRAN M., MAUREL D. (2006). Prolexbase : Un dictionnaire relationnel multilingue de noms propres. *TAL*, 47 : 3, 115-139.

TVEIT A., EDSBERG O., BROX RØST T., FAXVAAG A., NYTRØ Ø., NORDGÅRD T., THORSEN RANANG T., GRIMSMO A., (2004). Anonymization of General Practitioner Medical Records. *Second HelsIT Conference at the Healthcare Informatics,* Trondheim.

UZUNER O., LUO Y., SZOLOVITS P., (2007). *Evaluating the state-of-the-art in automatic de-identification*. J Am Med Inform Assoc 14, 550-63.

# Étude des risques de réidentification des patients à partir d'un corpus désidentifié de comptes-rendus cliniques en français

Cyril Grouin[1] Nicolas Griffon[2, 3] Aurélie Névéol[1]
(1) CNRS, LIMSI, UPR 3251, Campus universitaire d'Orsay, rue John von Neumann, 91405 Orsay
(2) INSERM, LIMICS, UMR_S 1142, 15 rue de l'École de Médecine, 75006 Paris
(3) CISMeF-TIBS-LITIS EA 4108, CHU de Rouen 76031 Rouen
prenom.nom@limsi.fr, prenom.nom@chu-rouen.fr

**Résumé.** La désidentification permet de préserver le secret médical lors de l'utilisation de documents cliniques pour faire avancer la recherche médicale. Cet article présente une évaluation des risques de réidentification des patients sur un corpus désidentifié de comptes-rendus cliniques en français. Les informations identifiantes sont marquées automatiquement dans le corpus, puis remplacées par des substituts plausibles. Les documents ainsi désidentifiés sont présentés à six évaluateurs avec une connaissance variable des documents et de la méthode de désidentification employée, afin qu'ils réidentifient les patients. La quantité d'informations identifiantes retrouvées semble liée à la familiarité des évaluateurs avec les documents et la méthode de désidentification. L'introduction de substituts géographiques de la même provenance que les documents originaux semble mieux préserver la confidentialité. Les informations retrouvées par les évaluateurs ne permettent pas de réidentifier les patients, sauf en cas d'accès privilégié au système d'information hospitalier de l'établissement d'origine des documents.

## Abstract.

**Chance of reidentification of patients from a de-identified corpus of clinical records in French**
De-identification aims at preserving patient confidentiality while enabling the use of clinical documents for furthering medical research. Herein, we evaluate patient re-identification risks on a corpus of clinical documents in French. Personal Health Identifiers are automatically marked by a de-identification system applied to the corpus, followed by reintroduction of plausible surrogates. The resulting documents are shown to individuals with varying knowledge of the documents and de-identification method. The individuals are asked to re-identify the patients. The amount of information recovered increases with familiarity with the documents and/or de-identification method. Surrogate re-introduction with localization from the same (vs. different) geographical area as the original documents is more effective. The amount of information recovered was not sufficient to re-identify any of the patients, except when privileged access to the hospital health information system and several documents about the same patient were available.

**Mots-clés :** Désidentification, réidentification, dossiers médicaux éléctroniques, vie privée.

**Keywords:** De-identification, Re-identification, Electronic Health Records, Privacy.

## 1 Introduction

Les recherches fondées sur des données cliniques supposent l'obtention du consentement des patients concernés. En France, les règles et recommandations en matière de respect de la vie privée impliquent que, dans les cas où il est impossible d'obtenir le consentement des patients (patient décédé, difficulté d'identifier les ayant-droits, etc.), les comptes-rendus cliniques doivent être anonymisés pour pouvoir être utilisés à des fins de recherche, en dehors du parcours de soin classique. La désidentification consiste à masquer les informations identifiantes relatives au patient, de telle sorte qu'il n'est plus possible de retrouver l'identité du patient (réidentification) sur la base des informations qui auront été laissées en clair. Les méthodes de désidentification automatique sont souvent évaluées sur leur capacité à identifier des éléments relevant de catégories prédéfinies depuis des comptes-rendus cliniques (Meystre *et al.*, 2010). Aux États-Unis, dix-huit catégories ont été définies dans le cadre de la loi HIPAA [1] de 1996 (US Department of Health Human Services, 1996). En

1. HIPAA : Health Insurance Portability and Accountability Act

l'absence d'une loi Européenne équivalente, nous utilisons ce cadre juridique pour désidentifier les documents cliniques français.

Évaluer le risque de réidentification des patients à partir de documents désidentifiés est une tâche complexe, dans la mesure où la combinaison d'éléments d'information en apparence inoffensifs peut néanmoins remettre en cause le respect de la vie privée du patient (Benitez & Malin, 2010; Barbaro & Zeller Jr, 2006; Grouin, 2013). La création de corpus de comptes-rendus cliniques réalistes a été réalisée avec succès (Neamatullah *et al.*, 2008) en enchaînant deux systèmes, un premier système d'identification des éléments à désidentifier dans les comptes-rendus cliniques suivi d'un deuxième système de remplacement des informations précédemment identifiées par des éléments fictifs plausibles. Le résultat produit un corpus valable sur les plans cliniques et linguistiques.

Alors que l'impact de la désidentification sur des traitements ultérieurs (étiquetage en parties du discours, extraction d'information, repérage d'entités nommées, etc.) a été étudié sur un corpus de comptes-rendus cliniques (Deléger *et al.*, 2013; Meystre *et al.*, 2014b), il n'existe que peu d'études sur l'impact réel concernant la vie privée des patients. Il a récemment été démontré que les médecins ne sont plus capables de ré-identifier les patients qu'ils ont récemment soignés au-delà d'un délai de trois mois, lorsqu'ils se fondent sur les comptes-rendus cliniques (Meystre *et al.*, 2014a). Il est cependant nécessaire d'aller au-delà de ces premières expériences, pour évaluer la nature et la possibilité des risques de réidentification des patients.

Dans cet article, nous présentons les expériences que nous avons menées en matière d'évaluation des risques de réidentification des patients par des humains, à partir de comptes-rendus cliniques rédigés en français et désidentifiés automatiquement. Les personnes impliquées dans cette étude (chercheurs en informatiques et médecin) présentent différents niveaux de connaissances, tant du point de vue des comptes-rendus cliniques étudiés que de celui des méthodes utilisées pour désidentifier automatiquement les comptes-rendus. Nous avons réalisé nos expériences sur différents types de données (documents relatifs au même patient ou à des patients différents) en faisant également varier les méthodes de réintroduction d'informations fictives (en utilisant soit des informations géographiques similaires à celles d'origine, soit des informations différentes).

## 2 État de l'art

Toute mise à disposition de données contenant des informations personnelles implique de respecter la vie privée des personnes mentionnées dans les données. Lorsque des documents désidentifiés sont créés à partir de documents réels pour être mis à disposition à des fins de recherche, il est nécessaire d'évaluer les risques de non respect de la vie privée au regard des bénéfices attendus par les résultats de la recherche utilisant ces données.

Lors des premières mises à disposition de données aux États-Unis, une évaluation inadéquate des risques de non respect de la vie privée a conduit à des situations délicates qui ont impliqué des actions en justice (Barbaro & Zeller Jr, 2006). À la lumière de cette expérience, des précautions extrêmes sont désormais requises avant toute mise à disposition de données dites « sensibles ». Le cas des données médicales, notamment celles contenues dans les comptes-rendus hospitaliers, nécessite une attention particulière. Fournir des données médicales qui ne respecteraient pas la vie privée du patient constituerait une violation du secret médical garanti dans le serment d'Hippocrate, le code de déontologie des médecins et le code pénal.

La base de données MIMIC II [2] (Saeed *et al.*, 2002, 2011; Lee *et al.*, 2011) constitue un exemple de réussite en matière de partage de données cliniques à grande échelle [3] respectant la vie privée des patients. En plus d'appliquer une méthode de désidentification performante, les concepteurs de la base de données ont mis en place un accord d'utilisation des données qui impose aux futurs utilisateurs d'être informés de la nature sensible de ces données et de veiller au respect de la protection de la vie privée au cas où ils identifieraient, dans les données fournies, des éléments nominatifs. À notre connaissance, il s'agit de la seule base de données de cette taille disponible pour la recherche clinique ou le traitement automatique des langues. De plus petits jeux de données cliniques ont également été distribués dans des conditions similaires à celles de la base MIMIC lors des campagnes d'évaluations, telle que les campagnes i2b2 dont la première édition en 2006 a porté sur la problématique de la désidentification des comptes-rendus cliniques (Uzuner *et al.*, 2007).

Nous estimons que l'étude des risques de réidentification des patients à partir de données désidentifiées et porteuses d'informations réalistes permet de mieux comprendre l'équilibre bénéfice/risque préalablement à la mise à disposition de

2. MIMIC : Multiparameter Intelligent Monitoring in Intensive Care
3. La version 3 de MIMIC comporte 23 180 dossiers.

données « sensibles ». D'autre part, nous pensons que ce type d'étude peut également contribuer à améliorer la conception et l'évaluation des systèmes de désidentification automatique, jusqu'à présent uniquement évalués en termes quantitatifs.

# 3 Matériel et méthodes

## 3.1 Corpus

Nous précisons que le corpus utilisé dans cette étude a reçu une autorisation de la CNIL [4], pour réaliser des recherches en matière de recherche d'information depuis un volume important de comptes-rendus électroniques patients. Dans notre étude, nous avons ciblé douze types d'informations à désidentifier, relatifs aux patients, aux parents des patients, et aux professionnels de santé : *prénoms, noms, initiales, adresses postales, villes, codes postaux, numéros de téléphone et de télécopie, adresses électroniques, noms d'hôpitaux, identifiants* (numéros de sécurité sociale ou numéros de série d'appareillage médical), et *dates* (y compris les dates de naissance) [5]. En matière de documents, nous avons sélectionné les trois types de documents les plus fréquents dans le corpus global d'où proviennent les données : compte-rendu hospitalier, compte-rendus d'acte, correspondance.

La désidentification a été réalisée au moyen d'approches à base d'apprentissage statistique, en appliquant l'outil MEDINA (conçu par l'utilisateur dénommé « Dev 2 » dans la suite de nos expériences). Nous avons construit un modèle statistique CRF (champs aléatoires conditionnels (Lafferty *et al.*, 2001))) adapté aux données à désidentifier sur la base d'un corpus de 100 documents annotés et vérifiés par des humains. L'approche utilisée, réalisée par les utilisateurs « Dev 1 » et « Dev 2 », a été décrite dans (Grouin & Névéol, 2014).

Dans la suite de cet article, nous appelons « données identifiantes réelles » l'ensemble des données présentes dans les documents d'origine qui correspondent aux douze types d'informations précédemment listés. Les « données réelles résiduelles » constituent un sous-ensemble. Elles correspondent aux données d'origine qui n'ont pas été identifiées automatiquement par notre système de désidentification MEDINA et qui n'ont donc pas fait l'objet d'un remplacement par des données fictives réalistes.

### 3.1.1 Critères d'inclusion

Grâce à l'outil MEDINA, nous avons volontairement constitué un corpus de travail comportant des documents pour lesquels il est fortement probable que des informations personnelles n'aient pas été identifiées, et restent visibles dans les documents transformés. Ainsi, notre étude permet d'évaluer les risques de ré-identification dans un contexte où ce risque peut être considéré comme élevé. Nous avons extrait du corpus désidentifié 60 documents pour lesquels nous savons que des informations ont échappé à l'outil de désidentification. Cette extraction repose sur différents critères jugés difficiles à appréhender pour un outil de désidentification automatique, critères que nous avons établis suite à une analyse des erreurs du système de désidentification :

- **noms, prénoms :** l'outil échoue à identifier des noms complexes, ou des portions de noms complexes, qui intègrent des traits-d'union ou des espaces (*Dorothy Jane, Watterman-Smith*) ;
- **informations de contact :** l'outil échoue également à identifier les informations de contact qui apparaissent dans le contenu même du document (c.-à-d., en dehors des entêtes et pieds-de-page), quand bien même ces informations sont introduites par des déclencheurs (*domicilié, personne de confiance*) ;
- **dates :** l'outil ne permet pas de faire la différence entre les dates liées au patient et les dates plus générales mentionnées dans le document (dates de procédure légale ou de changement de numérotation téléphonique). Appliquer le processus d'antédatation sur ces dates générales (remplacement des dates par des dates situées dans le passé en conservant le même écart pour toutes les dates d'un dossier) peut compromettre le processus global de désidentification puisque l'écart appliqué sur l'ensemble des dates du document permet alors de retrouver les dates d'origine.

4. CNIL : Commission nationale de l'informatique et des libertés `http:www.cnil.fr`

5. Ces douze catégories correspondent majoritairement aux catégories d'origine du HIPAA (US Department of Health Human Services, 1996). Certaines catégories définies par le HIPAA ne se retrouvent pas dans les documents français *(numéros de permis de conduire/carte d'identité, identifiants des véhicules, adresses Internet et adresses IP, identifiants biométriques, photographie)*. D'autre part, nous avons pris en compte les catégories *initiales* (des médecins) et *noms d'hôpitaux*, hors HIPAA, car nous considérons que ces informations, si elles ne permettent pas une réidentification directe, permettent néanmoins de réduire la population aux seuls patients traités par un ensemble donné d'hôpitaux.

### 3.1.2 Hypothèses d'évaluation du risque de réidentification

Les informations personnelles identifiées automatiquement ont été remplacées par des données fictives réalistes. Parce que l'outil de repérage automatique des informations à désidentifier aura manqué certaines de ces informations, en l'absence de vérification humaine, il reste donc dans le corpus désidentifié des données identifiantes provenant des documents d'origine. L'idée sous-jacente dans la réintroduction de données fictives réalistes repose sur le principe « *caché au vu de tous* ». Nous émettons l'hypothèse que les données identifiantes d'origines seront moins visibles si elles figurent au milieu d'autres données fictives réalistes. Le module de remplacement des données identifiantes par des données fictives a été mis au point par l'un des auteurs (« Dev 2 ») et étendu et adapté au corpus par un autre auteur (« Dev 1 »).

Nous évaluons le risque de réidentification en tenant compte de deux situations différentes, toutes deux pertinentes pour les besoins de la recherche médicale réalisée à partir de comptes-rendus désidentifiés.

**Restriction à un ou plusieurs patients** Selon l'objectif médical poursuivi dans une étude, il peut être nécessaire d'utiliser un corpus de documents relatifs au même patient (pour étudier la chronologie d'un patient ou l'apparition et l'évolution d'une maladie), par opposition à des documents sélectionnés aléatoirement et appartenants à différents patients. Nous émettons l'hypothèse que le risque de réidentification est plus élevé sur un corpus de documents relatifs à un même patient, dans la mesure où le corpus fournit plus d'informations sur le patient, et qu'il offre également la possibilité de croiser des informations entre documents.

**Origine géographique des documents** L'outil de réintroduction de données fictives repose sur des listes préétablies pour chaque catégorie d'information à désidentifier. Il est possible de configurer l'outil pour restreindre la zone géographique à un département lors du remplacement des codes postaux, villes et hôpitaux. Nous envisageons deux expériences, l'une réintroduit des données du même département que dans les données d'origine, l'autre réintroduit des données géographiques d'un autre département. Nous émettons l'hypothèse qu'il est plus complexe d'identifier des données réelles identifiantes au milieu de données fictives issues du même département.

### 3.1.3 Organisation du corpus

En raison de ces différentes configurations, nous avons divisé le corpus en quatre parties :

1. Quinze documents relatifs au même patient, avec réintroduction de données géographiques du même département ;
2. Quinze documents relatifs au même patient, avec réintroduction de données géographiques d'un autre département que celui d'origine ;
3. Quinze documents relatifs à différents patients, avec réintroduction de données géographiques du même département ;
4. Quinze documents relatifs à différents patients, avec réintroduction de données géographiques d'un autre département que celui d'origine.

Nous avons collecté les documents relatifs au même patient de la manière suivante :
– parmi tous les documents d'un patient, nous sélectionnons aléatoirement les documents correspondant à au moins l'un des trois critères de désidentification jugés difficiles (voir section 3.1.1) ;
– puis nous sélectionnons aléatoirement différents documents parmi les précédents documents jusqu'à atteindre le nombre souhaité (15 documents).
Pour le sous-corpus de documents correspondant à plusieurs patients, nous avons appliqué la sélection suivante :
– sélection de trois documents pour chacun des trois critères difficiles ;
– sélection aléatoire des autres documents, qu'ils contiennent ou non des critères difficiles à traiter.
La sub-division entre département identique et département différent de celui d'origine se fait après avoir constitué ces sous-corpus de documents, en substituant, ou non, le numéro de département présent dans les codes postaux du document.

Nous avons réalisé cette sélection aléatoire sans intervention humaine, de telle sorte que les auteurs de l'étude qui ont participé à la phase d'annotation (« Dev 1 » et « Dev 2 ») connaissaient les critères d'inclusion, mais ne savaient quelles informations dans les documents extraits relevaient de cette sélection. Le schéma 1 résume le processus suivi pour constituer les quatre sous-ensembles de corpus utilisés dans le cadre de cette étude.

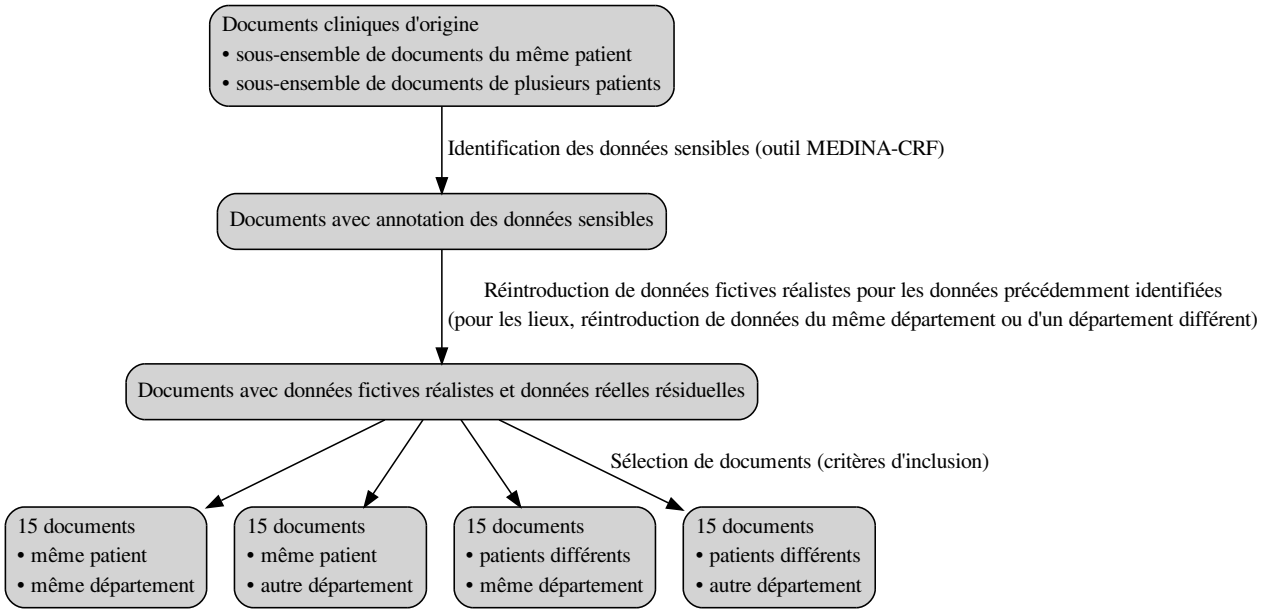


FIGURE 1 – Processus suivi pour constituer les corpus utilisés dans cette étude

#### 3.1.4 Constitution des corpus de références

Les 60 documents du corpus correspondent tous au résultat de l'identification des informations identifiantes par MEDINA avec substitution de ces informations par des informations fictives réalistes non-identifiantes. C'est sur les documents issus de ce traitement que les évaluateurs ont effectué leur travail de detection des données identifiantes originales (non substituées).

Après l'évaluation, deux des auteurs ont examiné l'ensemble des documents du corpus, dans leur version originale et leur version substituée, afin d'identifier toutes les données identifiantes. Deux versions de référence du corpus ont été créées, de manière à évaluer : $(i)$ les performances du système de désidentification automatique (évaluation qui ne peut se faire qu'à partir de la version d'origine du corpus [6]), et $(ii)$ les performances des annotateurs humains sur leur capacité à détecter les données identifiantes réelles (évaluation réalisée sur la version du corpus après réintroduction de données fictives). Si les deux versions du corpus de référence se rapportent aux 60 mêmes documents, la principale différence repose sur le fait que des données fictives ont été introduites en remplacement des données identifiantes réelles pour permettre la diffusion du corpus auprès des annotateurs humains. Parce que les entités ne sont plus les mêmes (*Pierre Fontaine* devient *Paul Martin*) d'une part, et que les coordonnées de début et de fin des portions à traiter changent [7] d'autre part, il n'est plus possible d'utiliser le même corpus de référence pour réaliser les deux évaluations. Nous donnons dans le tableau 1 un exemple de différences entre les deux versions de référence du corpus.

| | Texte d'origine (données identifiantes réelles) | Texte avec substitutions (données fictives) |
|---|---|---|
| Texte | *Pierre Fontaine né le 03/05/1979* | *Paul Martin né le 01/09/1977* |
| Annotations de référence (offsets de début et de fin) | Prénom : Pierre (0-6)<br>Nom : Fontaine (7-15)<br>Date : 03/05/1979 (22-32) | Prénom : Paul (0-4)<br>Nom : Martin (5-10)<br>Date : 01/09/1977 (17-27) |
| Utilisation | Evaluation de l'outil de désidentification automatique MEDINA | Evaluation des annotations humaines |

TABLE 1 – Exemples d'annotations de référence sur les deux versions du corpus

6. La version d'origine du corpus contient des données identifiantes réelles.

7. En raison du remplacement des données identifiantes réelles du corpus d'origine par des données fictives, les coordonnées de début et de fin de chaque portion contenant des données à traiter ne sont plus les mêmes. En effet, les offsets de caractères (depuis le début du fichier) diffèrent entre les deux version en raison des différences de taille (nombre de caractères) entre données identifiantes d'origine (*Pierre*, 6 caractères) et données fictives réintroduites (*Paul*, 4 caractères). L'outil utilisé pour l'évaluation (BRATeval) étant sensible aux formes de surface et aux différences d'offsets de caractères, il est donc nécessaire de disposer de deux versions de référence des corpus traités.

**Données identifiantes réelles vs. données non-identifiantes fictives** Une première version identifie les informations identifiantes réelles que l'outil de désidentification a échoué à identifier, et qui n'ont donc pas été remplacées par des données fictives réalistes (non-identifiantes). Cette version a été obtenue par comparaison des documents avant et après réintroduction des données fictives. Cette version comporte la liste des données identifiantes non retrouvées par MEDINA, ainsi que quelques cas de données identifiantes retrouvées par MEDINA mais remplacées à l'identique. Elle est utilisée pour évaluer la capacité des annotateurs humains à identifier les données identifiantes réelles dans le corpus désidentifié.

**Données à désidentifier** La deuxième version de référence identifie toutes les données à désidentifier dans le corpus d'origine. Cette version a été obtenue par comparaison du corpus avant et après identification par MEDINA des données à désidentifier. Cette version comporte la liste exhaustive des données identifiantes contenues dans le corpus, et sert à évaluer les performances du système de désidentification MEDINA sur ce corpus.

### 3.2 Expériences de réidentification

**Protocole expérimental** Nous avons soumis le corpus à trois catégories d'expérimentateurs humains, ayant différents niveaux de connaissance des documents (donc des informations d'origine susceptibles d'être présentes) et du fonctionnement de l'outil de désidentification (et des limites du systèmes sur certaines catégories) : $(i)$ un médecin de l'hôpital ayant fourni les données d'origine (nommé « Médecin » dans les expériences), $(ii)$ deux chercheurs en informatique ayant conçu et adapté l'outil de désidentification (nommés « Dev 1 » et « Dev 2 »), et $(iii)$ trois autres chercheurs ayant des compétences en informatique ou en linguistique, sans connaissance particulière du corpus ou de l'outil de désidentification (nommés « Chercheur 1 » à « Chercheur 3 »).

**Consignes** Il a été demandé à chaque expérimentateur d'annoter, dans les documents désidentifiés avec données fictives et présence résiduelle de données identifiantes réelles, toutes les données identifiantes qu'il estimait réelles, donc susceptibles de réidentifier pleinement ou partiellement le patient. Les annotations ont été réalisées au moyen de l'interface d'annotation BRAT (Stenetorp *et al.*, 2012). Pour les annotateurs qui n'étaient familiers, ni avec les données d'origine, ni avec l'outil de désidentification (« Chercheur » 1 à 3), nous leur avons présenté le processus global de constitution du corpus (sections 3.1 et 3.1.4). Contrairement aux autres annotateurs, ces annotateurs n'ont pas eu connaissance du département géographique d'où proviennent les données.

Une fois ce travail d'annotation terminé, nous avons interrogé chaque expérimentateur sur les indices qu'il a estimé utiles pour identifier les données identifiantes réelles. Ces indices sont variables selon les annotateurs et concernent n'importe quel élément ou combinaison d'éléments du texte susceptible d'être réel et permettant une réidentification potentielle. Nous renseignons en section 5.2 de ces tentatives de réidentification et des indices utilisés par les personnes interrogées.

## 4 Résultats

En moyenne, les expérimentateurs ont mis 2 heures pour traiter l'ensemble du corpus. Le travail d'annotation a ensuite été évalué par rapport aux annotations de référence, mais également en termes d'accord inter-annotateurs (F-mesure), tant sur le corpus global que sur chacun des quatre sous-corpus.

### 4.1 Performances initiales du système de désidentification

Nous présentons dans le tableau 2 la distribution des données identifiantes (nombre total et nombre de données identifiantes réelles) pour chaque catégorie dans le corpus final de 60 documents. Environ 9,5 % des données sont des données identifiantes réelles, alors que 90,5 % constituent des données fictives qui ont été réintroduites. On observe que la présence résiduelle d'informations identifiantes réelles n'est pas la même selon les catégories d'information, et qu'elle concerne majoritairement les *initiales* (89,7 %) et les *identifiants* (80,0 %), deux catégories qui se rapportent aux médecins mentionnés dans les comptes-rendus. Il en est de même pour les *noms* (3,3 %) et *prénoms* (3,5 %) dont aucun élément ne concerne un patient. Les informations résiduelles des catégories *adresses* (51,7 %), *codes postaux* (17,9 %) et *villes* (25,5 %) renvoient majoritairement aux coordonnées d'hôpitaux, en lien avec les informations résiduelles réelles de la catégorie *hôpitaux* (14,5 %). La catégorie *dates* est particulière, car nous recensons comme "données identifiantes réelles"

les dates qui permettent une identification des dates originales. Il s'agit donc à la fois des dates médicales non substituées (la date d'un examen effectué sur un patient) et des dates administratives de notoriété publique substituées (la date de changement de numérotation téléphonique).

| | NOM | PRE | INI | ADR | VIL | CP | TEL | MAIL | ID | DATE | HOP | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Total** | 541 | 487 | 39 | 60 | 153 | 67 | 282 | 42 | 20 | 233 | 166 | 2090 |
| **Réelles** | 18 | 17 | 35 | 21 | 39 | 12 | 0 | 0 | 16 | 17 | 24 | 199 |
| | (3,3%) | (3,5%) | (89,7%) | (51,7%) | (25,5%) | (17,9%) | (0,0%) | (0,0%) | (80,0%) | (7,3%) | (14,5%) | (9,5%) |

TABLE 2 – Distribution des données identifiantes totales et réelles par catégorie (NOM=noms, PRE=prénoms, INI=initiales, ADR=adresses, VIL=villes, CP=codes postaux, TEL=téléphones, MAIL=e-mails, ID=identifiants, DATE=dates, HOP=hôpitaux) dans le corpus final

Nous rapportons dans le tableau 3 les performances détaillées de l'outil de désidentification MEDINA en termes d'appariements à l'identique avec les données de référence [8]. La performance globale est de 0,93 de F-mesure, pour une précision de 0,96 et un rappel de 0,90, ce qui classe l'outil parmi les systèmes état de l'art du domaine.

| Catégorie | Précision | Rappel | F-mesure |
|---|---|---|---|
| Noms | 0,97 | 0,95 | 0,96 |
| Prénoms | 0,98 | 0,96 | 0,97 |
| Initiales | 0,67 | 0,05 | 0,09 |
| Identifiants | 1,00 | 0,25 | 0,40 |
| Hôpitaux | 0,74 | 0,53 | 0,62 |
| Adresses | 0,98 | 0,82 | 0,89 |
| Codes postaux | 1,00 | 0,79 | 0,88 |
| Villes | 0,99 | 0,95 | 0,97 |
| Dates | 0,94 | 0,97 | 0,96 |
| E-mails | 1,00 | 1,00 | 1,00 |
| Téléphones | 0,99 | 1,00 | 0,99 |
| **Total (micro-moyenne)** | 0,96 | 0,90 | 0,93 |

TABLE 3 – Performance de l'outil MEDINA, utilisant un modèle CRF appris sur un corpus de 100 documents

## 4.2 Exemple de document annoté

Nous présentons un extrait document (figure 2) qui a été jugé difficile à désidentifier automatiquement pour le critère "informations de contact" (voir section 3.1.1), en raison de la présence du déclencheur *personne de confiance*. Alors que le numéro de téléphone du mari de la patiente a correctement été détecté par MEDINA, les autres informations relatives à la famille de la patiente n'ont pas été détectées. Sur cet exemple, nous soulignons en vert les données fictives réintroduites, et encadrons en violet les données identifiantes réelles (qui n'ont donc pas été identifiées ni substituées). Nous précisons que les documents présentés aux expérimentateurs n'étaient porteurs d'aucune indication.

Dans cet exemple, les données identifiantes réelles concernent le lieu de résidence des enfants de la patiente : « *2 à Marseille et 1 en Corse* ». Deux annotateurs (« Dev 1 » et « Dev 2 ») ont correctement identifié ces deux informations comme étant réelles, et un annotateur (« Chercheur 1 ») n'a identifié que l'information *Corse*. Sur la base de ses connaissances, l'expérimentateur « Dev 1 » a cru identifier le numéro de téléphone comme étant réel, alors que dans le cas présent il s'agissait bien d'une donnée fictive réaliste.

8. Nous précisons que des données identifiantes d'origine ont pu être correctement identifiées par l'outil de désidentification automatique MEDINA, mais qu'elles auront été substituées par elles-mêmes sous l'effet du tirage au hasard (par exemple, lorsque la substitution substitue un nom de ville « *MARSEILLE* » par lui-même avec une différence de casse typographique « *Marseille* »). Dans ce cas, nous considérons qu'il s'agit toujours d'une donnée identifiante réelle malgré la substitution. En conséquence, il peut exister un écart entre les chiffres présentés dans les tableaux 2 et 3.

nom Huet prénom Arnaud
Née le : date 05/08/1928
Mode de vie :
- Situation familiale : Mariée. 3 Enfants (2 à ville Marseille et 1 en pays Corse).
- Lieu de vie : appartement au 5ème étage avec ascenseur.
- Ancienne profession : secrétaire de direction.

Réseau de soutien :
- Nom personne de confiance : époux N° Tél. : téléphone 06 19 46 13 89
- Aides à domicile : Aide ménagère 3 heures par semaine en chèque emploi service pour le gros ménage.
- APA : non.
ANTECEDENTS :
- Prothèse de hanche.
- Pathologie pancréatique en date 1993.
CONDUITE PROPOSEE :
- Consultation de suivi dans 1 mois pour évaluer la tolérance au traitement (date 26 novembre 2006).

Dr prénom Daniel nom Lucas
Médecin attaché.

FIGURE 2 – Extrait du corpus étudié. Les données fictives réintroduites sont encadrées en vert, les données identifiantes réelles résiduelles sont encadrées en violet (dans cet exemple, elles ont été remplacées par des données fictives)

### 4.3 Performances individuelles

Le tableau 4 présente les résultats détaillés par catégorie, au niveau global (ligne 6) et par sous-corpus (lignes 2 à 5), pour chaque expérimentateur, classé en fonction des connaissances de chacun sur les données et la méthode de désidentification. Trois groupes peuvent être distingués, séparés par des doubles barres : $(i)$ connaissances avancées à la fois sur les données et sur l'outil de désidentification, $(ii)$ connaissances avancées, soit sur les données, soit sur l'outil de désidentification, et $(iii)$ peu de connaissances sur les données et l'outil.

| **Corpus** | **Dev 1** N - P - R - F | **Médecin** N - P - R - F | **Dev 2** N - P - R - F | **Chercheur 1** N - P - R - F | **Chercheur 2** N - P - R - F | **Chercheur 3** N - P - R - F |
|---|---|---|---|---|---|---|
| 1 | 34 .71 .33 .45 | 13 .62 .11 .19 | 285 .16 .64 .26 | 30 .47 .19 .27 | 0 .00 .00 .00 | 26 .00 .00 .00 |
| 2 | 35 .57 .54 .56 | 11 .64 .18 .29 | 59 .19 .30 .23 | 8 .50 .11 .18 | 66 .02 .03 .02 | 24 .00 .00 .00 |
| 3 | 31 .61 .40 .49 | 19 .47 .19 .27 | 28 .71 .43 .53 | 6 .33 .04 .08 | 0 .00 .00 .00 | 26 .00 .00 .00 |
| 4 | 42 .67 .50 .57 | 25 .76 .34 .47 | 41 .51 .38 .43 | 15 .80 .21 .34 | 43 .07 .05 .06 | 10 .00 .00 .00 |
| **total** | 135 .62 .41 .50 | 66 .61 .20 .30 | 408 .23 .46 .30 | 56 .54 .15 .23 | 109 .04 .02 .03 | 84 .00 .00 .00 |

TABLE 4 – Performances sur l'identification des données identifiantes réelles en termes de nombre de données annotées (N), précision (P), rappel (R) et F-mesure (F). Les sous-corpus sont décrits en section 3.1.3 : 1 = même patient, même département, 2 = même patient, autre département, 3 = patients différents, même département, 4 = patients différents, autre département

Le tableau 5 présente les accords inter-annotateurs sur l'identification des données identifiantes réelles, calculés en termes de F-mesure.

| | Dev 1 | Médecin | Dev 2 | Chercheur 1 | Chercheur 2 |
|---|---|---|---|---|---|
| **Médecin** | 0,32 | - | | | |
| **Dev 2** | 0,21 | 0,10 | - | | |
| **Chercheur 1** | 0,21 | 0,11 | 0,18 | - | |
| **Chercheur 2** | 0,00 | 0,00 | 0,00 | 0,01 | - |
| **Chercheur 3** | 0,01 | 0,01 | 0,03 | 0,01 | 0,00 |

TABLE 5 – Accords inter-annotateurs (F-mesure)

# 5 Discussion

## 5.1 Performances individuelles d'identification des données identifiantes réelles

Le tableau 4 montre que globalement, l'identification de données identifiantes réelles est faible, et que la capacité de l'expérimentateur à détecter ces données est fortement lié aux connaissances antérieures qu'il a des documents et/ou de l'outil de désidentification. La F-mesure la plus élevée est de 0,50 (expérimentateur cumulant connaissances fines sur le corpus et sur l'outil de désidentification), ce qui ne constitue pour autant pas une bonne performance. Alors que les chercheurs 1 à 3 n'avaient aucune connaissance des données et de l'outil, le chercheur 1 a néanmoins deviné l'origine géographique des documents une fois l'expérience terminée. Cependant, personne n'a été en mesure d'identifier plus d'informations spécifiques aux patients.

Le tableau 5 montre que l'accord inter-annotateur le plus élevé est observé entre l'expérimentateur ayant obtenu les meilleurs résultats (« Dev 1 ») et le médecin, autrement dit, les deux expérimentateurs ayant des connaissances fines des données d'origine. L'accord est cependant faible à 0,33 de F-mesure, ce qui ne permet pas de conclure à un accord entre annotateurs (Artstein & Poesio, 2008). Ainsi, même en ayant des connaissances précises des données et/ou de l'outil de désidentification, deux humains ne parviennent pas à se mettre d'accord sur ce qui constitue une donnée identifiante réelle ou non. D'après ces résultats, la stratégie « *caché au vu de tous* » semble correctement fonctionner, et les données identifiantes réelles ne sont pas évidentes pour les différents expérimentateurs impliqués.

## 5.2 Tentatives de réidentification

Les outils disponibles pour tenter de réidentifier les patients se composent essentiellement des informations disponibles sur internet. L'un des expérimentateur (« Chercheur 1 ») a systématiquement vérifié les noms d'hôpitaux, de personnes et les adresses, en utilisant un moteur de recherche classique, ce qui lui a permis d'identifier l'hôpital d'où proviennent les données. Deux expérimentateurs (« Dev 1 » et « Médecin ») ont utilisé un annuaire inversé pour vérifier tous les numéros de téléphone et adresses qu'ils estimaient être réels. Les requêtes n'ont cependant renvoyé aucun résultat. Les autres annotateurs n'ont pas indiqué avoir utilisé d'autres sources d'informations que celles présentes dans les documents traités.

Un expérimentateur (« Médecin ») a eu accès au système d'information patient (SIP) de l'hôpital d'où sont extraites les données. Sur le sous-corpus mélangeant les documents relatifs à plusieurs patients, il ne lui a pas été possible de réidentifier le moindre patient, faute de pouvoir établir une requête valable dans le système [9] ; des essais de ré-identifications soutenus ont été effectués sur trois documents, et ont été abandonnés au bout de 30 minutes de recherches infructueuses. En revanche, pour les sous-corpus proposant plusieurs documents sur le même patient, sur la base d'un recoupement d'informations entre la date de séjour approximative et les codes d'actes médicaux trouvés dans les documents ou inferrés à partir de connaissances médicales (ces codes cliniques ne constituent pas des informations identifiantes), il a été possible au médecin de constituer une requête valide et de réidentifier correctement le patient. Pour les deux patients du corpus (sous-corpus 1 et 3), les expériences de réidentification dans le SIP ont cependant demandé 20 minutes pour un patient et 30 minutes pour le deuxième avant de pouvoir effectivement retrouver les identités d'origine.

L'outil le plus efficace pour réidentifier un patient se révèle donc être le système d'information de l'hôpital, à condition de disposer des droits d'accès à cet outil et de savoir l'utiliser. Le SIP est conçu de telle sorte que l'utilisateur doit fournir

9. Le système d'information patient est conçu de telle sorte qu'il n'est possible de l'interroger qu'au moyen d'une requête combinant suffisamment d'éléments pour correspondre à un patient enregistré dans la base. Une recherche en plein texte – comme dans un moteur de recherche classique – n'existe pas, car elle ne correspond pas à un besoin des médecins. Aucun outil déployé dans l'hôpital ne permet donc de faire une telle interrogation.

suffisamment d'informations en entrée sur un patient pour pouvoir accéder aux documents de ce patient. Dans notre étude, un seul document désidentifié avec données fictives ne permet pas de retrouver le patient d'origine. À l'inverse, lorsque plusieurs documents sont disponibles pour un même patient, le patient concerné peut être retrouvé. Mais dans ce cas, la réidentification du patient suppose : $(i)$ un accès au SIP, $(ii)$ une connaissance de la manière dont les documents sont codés et stockés dans le SIP, et $(iii)$ des connaissances médicales pour inférer le code des diagnostics à partir des documents, de manière à disposer d'informations complémentaires pour enrichir la requête dans le SIP.

### 5.3 Performances selon la configuration des expériences

Les résultats présentés dans la tableau 4 montrent que les performances globales sont généralement meilleures lorsque la réintroduction de données fictives se fait sur un département différent de celui des données d'origine (corpus 2 et 4). Cette observation révèle que le principe « *caché au vu de tous* » est conforté puisque les données identifiantes réelles résiduelles auront été « noyées » parmi les données fictives réintroduites sur le même département.

Un argument en défaveur de l'utilisation du même département que celui des données d'origine lors de la réintroduction de données fictives portait sur le fait que, potentiellement, il existe un risque non nul de tirer aléatoirement la même information que celle que l'on cherche à masquer *(un nom d'hôpital dans une liste d'hôpitaux, une ville parmi toutes les villes du département, etc.)*, ce qui réduit à néant les efforts de désidentification. Nous avons constaté ce phénomène dans le cas où l'information à réintroduire est dans une forme typographique différente de celle d'origine (par exemple, « *Beau-Mont* » vs. « *BEAUMONT* »).

Au regard des résultats présentés dans le tableau 4, il n'est pas évident de déterminer que l'identification est rendue plus simple lorsque le travail porte sur le même patient (corpus 1 et 2) que sur des patients différents (corpus 3 et 4). Cependant, la disponibilité de plusieurs documents pour un même patient facilite la réidentification du patient dans le SIP de l'hôpital.

Pour ce qui concerne les dates, l'objectif de la désidentification et de la pseudonymisation consiste à préserver la vie privée du patient tout en maintenant un caractère réaliste aux données contenues dans les documents. Dans le domaine médical, les préconisations en matière d'anonymisation précisent qu'« *un dispositif d'anonymisation doit permettre de suivre le dossier d'une même personne, non identifiable, dans la durée* » (Belleil, 2008).

### 5.4 Limites

L'une des limites de cette étude concerne la taille du corpus. Nous avons restreint la taille du corpus à 60 documents, de manière à proposer une durée de réalisation de l'expérience acceptable pour les différents expérimentateurs. Ce corpus se révèle comparable en termes de taille par rapport à celui utilisé par (Meystre *et al.*, 2014a), constitué de 85 documents. De fait, les configurations que nous souhaitions étudier nous ont conduit à diviser le corpus en sous-ensembles de 15 documents, ce qui permet uniquement de dégager des résultats indicatifs. Nous estimons que cette étude mériterait d'être poursuivie sur une plus grande échelle.

D'autre part, une autre catégorie importante de personnes capables de réidentifier les patients à partir du contenu des documents désidentifiés sont les patients eux-mêmes, ou les parents et proches des patients. Par exemple, une personne qui connait personnellement le patient présenté dans notre exemple (Figure 2) pourrait lire ce document et réaliser que les informations disponibles *(mère de 4 enfants, sans profession, avec une prothèse de hanche et un problème pancréatique dans le passé)* correspondent à l'une de ses connaissances. Nous n'avons cependant pas été en mesure de mettre en place un protocole expérimental adéquat pour mesurer ce risque, la démarche étant longue et coûteuse (recherche des patients, obtention de leur accord pour participer, explication des objectifs et de la procédure suivie, éventuels défraiements, etc.). Nous estimons que la possibilité de réidentifier un tel patient doit être similaire à celle qu'un médecin réidentifie un patient qu'il a soigné dans les trois derniers mois. Il a ainsi été démontré que les médecins ne sont pas capables de réidentifier leurs propres patients à partir de comptes-rendus cliniques désidentifiés au-delà de trois mois (Meystre *et al.*, 2014a).

## 6 Conclusion

Dans cet article, nous avons présenté les expériences que nous avons menées en matière d'identification, par des expérimentateurs humains, de données identifiantes réelles résiduelles dans un corpus de comptes-rendus cliniques désidentifiés

et porteur d'informations fictives réalistes. Nous avons également étendu ces expériences à des tentatives de réidentification des patients, compte-tenu des informations disponibles dans les documents ou des informations qu'il est possible d'inférer au moyen de connaissances médicales (ici, le codage des actes médicaux).

En dépit de l'absence de désidentification de certains éléments par notre outil de désidentification d'une part, et de la connaissance des faiblesses de l'outil par les développeurs d'autre part, jamais la protection de la vie privée des patients n'a été remise en cause sans disposer d'un accès privilégié au système d'information patient (SIP) de l'hôpital d'où sont issues les données, accès strictement réservé au personnel médical de l'établissement. Lorsqu'un accès au SIP est possible, les patients peuvent être réidentifiés par le biais d'un recoupement d'informations trouvées dans plusieurs documents et par la mobilisation de connaissances médicales sur le codage des actes médicaux. Le respect de la vie privée des patients semble néanmoins respecté lorsque n'est fourni qu'un seul document par patient, y compris pour le personnel médical disposant des accès au SIP.

Enfin, il est moins évident de retrouver des informations lorsque la réintroduction de données géographiques se fait à partir de données issues du même département que celui des données d'origine.

### Remerciements

Ce travail a reçu le soutien de l'Agence Nationale pour la Recherche (ANR) dans le cadre du projet CABeRneT (Compréhension Automatique de Textes Biomédicaux pour la Recherche Translationnelle) ANR-13-JS02-0009-01. Les auteurs remercient le service d'informatique medicale du CHU de Rouen pour l'accès au corpus LERUDI, ainsi que les annotateurs qui on participé à cette étude (Annick Choisier, François Morlane-Hondère, Kevin B. Cohen).

## Références

ARTSTEIN R. & POESIO M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, **34**(4), 555–96.

BARBARO M. & ZELLER JR T. (2006). A face is exposed for aol searcher no. 4417749. *The New York Times*.

BELLEIL A. (2008). *Référentiel AFCDP des dispositifs d'anonymisation*. Rapport interne, AFCDP.

BENITEZ K. & MALIN B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. *J Am Med Inform Assoc*, **17**(2), 169–77.

DELÉGER L., MOLNAR K., SAVOVA G., XIA F., LINGREN T., LI Q., MARSOLO K., JEGGA A., KAISER M., STOUTENBOROUGH L. & SOLTI I. (2013). Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *J Am Med Inform Assoc*, **20**(1), 84–94.

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

GROUIN C. & NÉVÉOL A. (2014). De-identification of clinical notes in french : towards a protocol for reference corpus developpement. *J Biomed Inform*, **46**(3), 506–515.

LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, p. 282–289, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

LEE J., SCOTT D. J., VILLARROEL M., CLIFFORD G. D., SAEED M. & MARK R. G. (2011). Open-access MIMIC-II database for intensive care research. In *Proc IEEE Eng Med Biol Soc*, p. 8315–8.

MEYSTRE S., SHEN S., HOFMANN D. & GUNDLAPALLI A. (2014a). Can physicians recognize their own patients in de-identified notes ? In *Stud Health Technol Inform*, volume 205, p. 778–82.

MEYSTRE S. M., FERRÁNDEZ O., FRIEDLIN F. J., SOUTH B. R., SHEN S. & SAMORE M. H. (2014b). Text de-identification for privacy protection : a study of its impact on clinical text information content. *J Biomed Inform*, **50**, 142–50.

MEYSTRE S. M., FRIEDLIN F. J., SOUTH B. R., SHEN S. & SAMORE M. H. (2010). Automatic de-identification of textual documents in the electronic health record : a review of recent research. *BMC Med Res Methodol*, **10**(70).

NEAMATULLAH I., DOUGLASS M. M., LEHMAN L.-W. H., REISNER A., VILLAROEL M., LONG W. J., SZOLOVITS P., MOODY G. B., MARK R. G. & CLIFFORD G. D. (2008). Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak*, **8**(32).

SAEED M., LIEU C., RABER G. & MARK R. G. (2002). MIMIC II : a massive temporal ICU patient database to support research in intelligent patient monitoring. *Comput Cardiol*, **29**, 641–4.

SAEED M., VILLAROEL M., REISNER A. T., CLIFFORD G., LEHMAN L.-W., MOODY G., HELDT T., KYAW T. H., MOODY B. & MARK R. G. (2011). Multiparameter intelligent monitoring in intensive care ii (MIMIC-II) : A public-access intensive care unit database. *Crit Care Med*, **39**(5), 952–60.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). BRAT : a web-based tool for NLP-assisted text annotation. In *Proc of EACL Demonstrations*, p. 102–7, Avignon, France : ACL.

US DEPARTMENT OF HEALTH HUMAN SERVICES S. . (1996). Health Insurance Portability and Accountability Act. http ://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/adminsimpregtext.pdf. §164.514.

UZUNER O., LUO Y. & SZOLOVITS P. (2007). Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc*, **14**(5), 550–63.

# Faire du TAL sur des données personnelles : un oxymore ?

Hugues de Mazancourt[1], Alain Couillault[2, 3], Gilles Adda[4, 5], Gaëlle Recourcé[6]
(1) Eptica, 95b rue de Bellevue, 92100 Boulogne-Billancourt
(2) L3i, Laboratoire Informatique, Image et Interaction, Université de La Rochelle, Avenue Michel Crépeau, 17042 La Rochelle, France
(3) APROLAB, Aproged, 43, rue Beaubourg, 75003 Paris
(4) Spoken Language Processing group, LIMSI-CNRS, Orsay (5) IMMI-CNRS, Orsay
(6) Kwaga Lab, Kwaga, 15, rue Jean Baptiste Berlier, 75013 Paris
hugues.de-mazancourt@eptica.com, alain.couillault@aproged.org, gadda@limsi.fr, recource@kwaga.com

**Résumé.** Le présent travail s'inscrit dans le cadre de la version 2 de la Charte Ethique et Big Data (Couillault & Fort, 2013). Il présente les difficultés inhérentes à l'application de techniques de TAL sur des données à caractère personnel, en même temps que la nécessité d'un tel travail.

## Abstract.

**Is NLP of personal data an oxymoron ?**
We present the work in progress for the version 2 of the Big Data Charter and present some aspects of the use of pesonal data for an industrial system embedding NLP technology.

**Mots-clés :** Données privées, Big Data, Ethique.

**Keywords:** Privacy, Big Data, Ethics.

## 1 Introduction

Notre réflexion, dans le cadre de la révision de la charte Ethique et Big Data, se centre sur l'oxymore qu'est l'impossible nécessité de réaliser des travaux de TAL sur des données à caractère personnel. Elle vise à généraliser et aller au delà des solutions mises en œuvre. Après une présentation de la Charte et du cadre légal de ces données, nous présentons quelques solutions mises en œuvre pour effectuer de tels travaux, en soulignant leurs limites. Nous montrerons que la Charte Ethique et Big Data, si elle ne résout pas tous les problèmes, fournit une méthodologique pour aborder ces traitements.

## 2 La charte Ethique et Big Data

La charte Ethique et Big Data (Couillault & Fort, 2013) a pour objectif de fournir un outil de documentation des jeux de données - et en particulier les ressources langagières -, afin d'en assurer la traçabilité et la transparence. Son principe de base est celui d'un questionnaire **déclaratif**, l'objet étant de fournir à l'utilisateur de la donnée, quelle qu'elle soit, des informations sur la façon dont elle a été produite et la licence attachée. Elle n'assure donc pas l'éthique mais fournit suffisamment d'informations à l'utilisateur pour qu'il puisse décider du degré d'éthique des données et de l'utilisation qu'il en fait.

### 2.1 Structure de la Charte

La Charte se présente comme un formulaire à remplir pour décrire une donnée, quelle qu'elle soit. Elle comprend une une première section de description des données fournies, puis s'organise autour des trois axes suivants via des séries de questions précises :

— traçabilité des travaux effectués sur les données (transformations) et des acteurs impliqués,
— explicitation de la licence d'usage,
— description des éventuelles législations et contraintes spécifiques

La nécessité d'une Charte pour les données a été clairement exposée par ses initiateurs (Couillault & Fort, 2013). La première version de la Charte a d'ores et déjà été adoptée par de nombreuses associations dont Cap Digital. Ainsi, de nombreux projets proposés au financement public (FUI et autres Appels à Projets gérés par BPIfrance) mentionnent la Charte dans leur description technique.

### 2.2 Motivations d'une nouvelle version

L'objet de la version 2 de la Charte est double. Il consiste en premier lieu à en faire un objet *actionnable*, c'est-à-dire un formulaire informatisé qui puisse être rempli, consulté en ligne et directement accessible avec les données qu'il décrit. Ainsi, l'objectif de traçabilité est-il complet, puisqu'on dispose du lien direct entre la Charte et la Donnée. Cette nouvelle version est également l'occasion de renforcer la Charte sur un certain nombre d'aspects. Il s'agit notamment de la confronter aux multiples initiatives assimilées en France ou en Europe. Parmi ces travaux, citons les *ethics guidelines* du programme H2020 (European Commission, 2014) ou les travaux dans le domaine de la santé (plus précisément dans les groupe d'intérêt TIC et Santé des Pôles de Compétitivité). Cette confrontation est l'objet d'un groupe de travail qui se réunit mensuellement, de février à juin 2015. Dernier point, on a parfois assimilé cette Charte à une utilisation uniquement liée aux données linguistiques, en raison de sa genèse. En effet, la Charte trouve son origine dans une réflexion sur les corpus constitués en TAL avec des outils de crowdsourcing peu respectueux du droit du travail (voir (Fort *et al.*, 2014)). La validation de la Charte sur ces différents travaux va définitivement invalider cette critique.

Un dernier objectif de cette nouvelle version est d'aborder explicitement le sujet des données personnelles, sujet qui était traité dans la première version par simple renvoi aux dispositions de la CNIL.

## 3 Le TAL et les données personnelles

La question des données personnelles agite de façon grandissante le monde du Traitement Automatique des Langues : le TAL a en effet besoin de corpus et un nombre croissant de ces corpus contient de telles données (de façon directe ou indirecte), entre autres avec l'abondance du contenu généré par les utilisateurs sur les forums ou réseaux sociaux. Ce contenu, riche en potentiel d'applications et en phénomènes linguistiques (et qui intéresse de plus en plus l'industrie) recèle des informations personnelles, privées, voire intimes, en particulier dans le domaine de la santé. Ces informations sont publiées par leurs auteurs sans grand souci de leur utilisation au delà d'une publication dans un forum à un instant donné. Faire du traitement automatique sur ces données est tentant, aisé, mais, comme le soulignent (Boyd & Crawford, 2011), l'accessibilité de la donnée ne rend pas automatiquement son utilisation éthique.

### 3.1 Les données personnelles

La CNIL définit une donnée personnelle comme suit [1] :

> Constitue une donnée à caractère personnel toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres. Pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens en vue de permettre son identification dont dispose ou auxquels peut avoir accès le responsable du traitement ou toute autre personne.

1. http://www.cnil.fr/documentation/textes-fondateurs/loi78-17/

Des discussions sont toujours en cours à la CNIL pour décider si, dans les faits, une donnée est ou non personnelle (cf. l'adresse IP qui a été intégrée récemment à cette définition), preuve s'il en est que le contour de cette notion est moins simple qu'il n'y paraît. A cette définition, on peut ajouter celle (toujours selon la CNIL) de **données sensibles** : *Les données sensibles sont celles qui font apparaître, directement ou indirectement, les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses ou l'appartenance syndicale des personnes, ou sont relatives à la santé ou à la vie sexuelle de celles-ci*. La collecte ou le traitement de données sensibles est interdit, sauf consentement explicite de l'intéressé (ou jugement spécifique impliquant une autorisation de la CNIL). D'une manière générale, le traitement de données personnelles (même non sensibles) est très strictement encadré par la loi dite *Informatique et Libertés* et ne doit, entre autres, *pas être réalisé pour d'autres finalités* que celles qui ont permis la collecte, sauf consentement éclairé de l'auteur [2]. Il est donc a priori illégal d'effectuer des traitements d'analyse sur un corpus collecté sur le web (ou pire, dans une boîte mail) pour peu qu'une information non-structurée [3] (un texte, un enregistrement) permette d'identifier un individu réel.

## 3.2 La question de l'anonymat

La première parade pour éviter cette identification est d'anonymiser les textes : plus de référence nominative [4], plus d'identification. De façon assez paradoxale, on demande au TAL d'outiller l'anonymisation, la reconnaissance d'entités nommées étant perçue comme la première étape (mais certes pas la seule, voir notamment (Amblard *et al.*, 2014) ou (Medlock, 2006)) pour détecter les éléments nominatifs dans un texte.

Une fois ces éléments identifiés, une solution populaire est la pseudonymisation, qui consiste à remplacer une entité par une autre entité de même type (un nom par un autre nom, un prénom par un autre prénom, etc.). Elle a le mérite de produire des contenus similaires à ceux d'origine, sur lesquels il sera possible d'entraîner des automates qui pourront ensuite être appliqués à de "vrais" textes. La pseudonymisation, permise par l'altération d'entités nommées en des éléments non-signifiants constitue ainsi pour (De Mazancourt *et al.*, 2014) un des niveaux d'anonymisation qui se situe, dans le cas des courriels entre l'anonymisation des méta-données et l'anonymisation dite *vraie*, définie comme l'impossibilité de restituer l'identité de l'auteur.

## 3.3 Impossibilité technique d'une anonymisation vraie

De nombreux chercheurs ont montré les limites d'une telle approche. Par exemple (Eshkol-Taravella *et al.*, 2014), démontre, lors du traitement du corpus oral ESLO (recueil d'interviews à Orléans en 1968), la présence d'éléments d'identification directs (*mon père a fondé le plus grand cabinet d'ophtalmologiste de la ville*) ou non directs (*le locuteur est patron de café au moment de l'enregistrement et il travaillait auparavant dans l'aviation militaire*) qui peuvent permettre, avec une connaissance raisonnable du contexte, de *réidentifier* les individus réels.

Récemment, dans un domaine proche, une étude (de Montjoye *et al.*, 2013) a montré qu'il suffisait de 4 points géolocalisés pour identifier de façon unique 95% des utilisateurs dans une base de plus d'un 1,5 million d'individus à partir des traces de connexion de leurs téléphones portables sur 6 mois. Sans aller toutefois jusqu'à la ré-identification il démontre qu'on peut lier de façon presque certaine une connexion isolée à tout le parcours du téléphone et donc de son utilisateur. De façon similaire, les outils de TAL, appliqués à des corpus suffisamment vastes, comme le sont ceux qui sont à disposition actuellement sur les forums ou réseaux sociaux, vont permettre d'établir des liens, de collecter un certain nombre d'indices sur les individus a priori anonymes. Ils vont permettre de déduire des textes que l'utilisateur X a une voiture, qu'il est allé l'année dernière aux Canaries et qu'il a acheté un smartphone de marque Y, simplement en analysant les "traces" qu'il laisse sur le Net. Ce sont ces corrélations dont sont friands les professionnels du marketing. Ces corrélations extraites des textes sont autant d'indices pour la réidentification. En faisant le parallèle avec les traces laissées par les téléphones portables, on imagine aisément qu'il est possible de collecter suffisamment d'indices à partir de textes pour connecter une portion significative de verbatims produits sur un réseau social, par exemple, même couverts par l'anonymat formel (masquage des identifants de toute sorte). Et dès lors, il suffira une fuite (par exemple) pour que ce parcours dans le réseau se mue en une masse d'informations personnelles et intimes sur des personnes réelles.

2. On retrouve cette notion de *consentement libre et éclairé* dans la relation soignant-malade, ainsi que stipulé dans l'Art. 16-3 du Code Civil
3. nous éliminons ici les méta-données structurées accompagnant les textes, pour lesquelles une étude linguistique n'est pas pertinente
4. au sens large, incluant tous les types d'identifants

# 4 Le cas du projet ODISAE

Si le TAL appliqué à des données à caractère personnel porte en soi le germe d'une violation potentielle de la vie privée, le linguiste ne doit pourtant pas s'abstenir de ce type de travail. En effet, l'étude de corpus de données à caractère personnel peut être particulièrement enrichissant sans que l'objectif soit de porter atteinte à la vie privée, même si elle peut en devenir un effet de bord. C'est le cas par exemple du projet ODISAE.

## 4.1 Présentation du projet

Le projet ODISAE, co-financé par BPIfrance et la Région Ile de France, labellisé par les Pôles de compétivité Cap Digital et Images et Réseaux dans le cadre du FUI-17 (Fonds Unifié Interministériel) réunit huit PME partenaires et un universitaire, le LINA. Le chef de file de ce projet est la société Eptica, éditeur de logiciel spécialisé dans la relation client. Les partenaires sont des éditeurs dans des domaines proches ou connexes (Cantoche, Kwaga, Jamespot, TokyWoky) ou organismes ayant un rôle de valideur de par une activité dans ce domaine (Aproged, Centre Départemental du Tourime de l'Aube et INSEE). L'objectif du projet est de fournir des outils logiciels innovants pour un centre de contact client en analysant le contenu des échanges réalisés entre le centre de contact et les utilisateurs. Le projet se focalise sur les échanges écrits (mail, chat, etc.). Il s'agit de considérer ces échanges comme un dialogue entre l'utilisateur et la marque (au sens large) et non pas comme une succession de messages déconnectés les uns des autres. On va par exemple tâcher de déclencher, avant qu'il ne soit trop tard, des actions lorsqu'une conversation se "passe mal" ou lorsque le client menace de changer de fournisseur (détection d'attrition). On cherche également à comprendre le degré d'adéquation de la FAQ, mise à disposition de ceux qui répondent aux clients, avec les problématiques qu'ils expriment au long de cet échange.

La première difficulté, pour un tel projet, est la faiblesse de l'état de l'art sur le sujet, faiblesse qui s'explique très simplement par un manque de corpus utilisables pour l'étude linguistique. En effet, le seul corpus d'e-mails à la disposition des chercheurs est le corpus Enron ((Klimt & Yang, 2004)) qui souffre de nombre de défauts pour une utilisation dans ce contexte, à commencer par son âge et ses conditions de réalisation. Il s'agit d'échanges datant de plus de dix ans, essentiellement entre collègues d'une même société (Enron). Or, en dix ans, l'usage de l'e-mail a considérablement changé. De plus, on ne s'adresse pas à ses collègues de la même façon qu'on s'adresse à une marque pour réclamer un remboursement.

Cette pénurie de corpus s'explique par le fait qu'un échange d'e-mails est une donnée éminemment personnelle, couverte de plus par des droits multiples comme le droit d'auteur et le droit à la correspondance privée. Dans le cadre du projet, la solution apportée pour permettre l'étude linguistique est décrite dans (De Mazancourt *et al.*, 2014). Elle situe sur deux plans :

1. sur un plan technique, les partenaires valideurs fournissent leurs corpus au consortium suite à une pseudonymisation effectuée dans les locaux du fournisseur de données. Aucune donnée *authentique* ne sort de chez ce partenaire, seul est fourni un corpus qui "ressemble" au corpus initial mais dont tous les noms et identifiants ont été remplacés par des entités similaires ;
2. sur un plan juridique, tous les membres du consortium sont soumis à un accord de confidentialité strict, empêchant la diffusion des corpus (ainsi pseudonymisés) en dehors du consortium.

Une phase d'anonymisation *vraie*, qui aurait impliqué la réécriture manuelle des e-mails afin d'en masquer toute information directe ou indirecte n'a pas été retenue lors du financement du projet. Elle aurait été particulièrement coûteuse.

## 4.2 Limites de la solution

Si la solution mise en place dans ODISAE permet au consortium de travailler et de produire les résultats attendus d'un point de vue opérationnel, elle se heurte à deux critiques du point scientifique. D'une part, le projet n'aura pas fait progresser la communauté sur l'aspect de la disponibilité d'un corpus d'e-mails puisque les données ne peuvent être diffusées à l'extérieur du consortium. D'autre part, les données restant confidentielles, il n'est pas possible de réfuter scientifiquement les conclusions qui auront pu être tirées de cette étude par les universitaires, la matière première permettant de valider ou d'invalider ces conclusions n'étant pas disponible.

Cette situation ne se limite pas aux mails. Nous avons évoqué précédemment le cas des données collectées sur les réseaux sociaux et on peut aussi citer les travaux sur des compte-rendus d'entretiens médicaux ((Amblard & Fort, 2014)) et nombre d'autres domaines encore pour lesquels les travaux en TAL ne peuvent être scientifiquement étayés par la publication des

corpus d'étude. Il n'est pas question de mettre en cause l'interdiction de cette diffusion mais bien de pointer la difficulté scientifique que cet interdit implique.

Cette mise en porte-à-faux des scientifiques porte peut-être en elle une piste de solution. En effet, si l'on excepte les applications de renseignement, les industriels ne sont pas a priori demandeurs de corpus de données à caractère nominatif. Ils sont demandeurs d'études linguistiques, de modélisations, voire de modèles effectifs pour des systèmes d'apprentissage. Leur tâche est d'industrialiser ces modélisations pour les mettre en œuvre dans des systèmes opérationnels. Or pour que les modèles puissent voir le jour, il faudrait que les scientifiques (et a priori seulement eux) aient accès aux corpus permettant de fabriquer ces modèles. Mais force est de constater qu'aujourd'hui, la fabrication de tels modèles dans un cadre industriel ne peut s'effectuer qu'au prix d'une activité scientifique affaiblie, voire sans étude scientifique du tout.

## 5 Résoudre le paradoxe

Pour faire du TAL avec des données personnelles, il faut que plusieurs questions soient résolues, en premier lieu délimiter ce qu'est une donnée personnelle. Une définition juridique existe, mais sa portée éthique n'est pas toujours claire et est de plus variable pour chacun. Le deuxième pas est d'outiller l'anonymisation vraie, qui est un objectif asymptotique et mouvant car les techniques de réidentification vont nécessairement évoluer. Enfin, fournir un cadre juridique à leur utilisation. La solution n'est pas uniquement technique, pas plus qu'elle n'est uniquement légale ou éthique, mais une combinaison de tous ces axes :

1. la technique doit assurer qu'un niveau suffisant de traitement a été réalisé sur les données pour rendre la réidentification la plus complexe possible ;
2. la licence associée aux données doit explicitement indiquer ce qui est possible ou non d'en faire afin de placer son utilisation dans un cadre juridique précis ;
3. enfin, la transparence du processus depuis la collecte des données jusqu'à leur fourniture doit permettre d'en assurer l'aspect éthique.

La Charte Ethique et Big Data fournit un cadre méthodologique permettant de décrire ces trois aspects, de fournir un descriptif qui accompagne la donnée tout au long de sa vie. On pourra alors, si le droit, les techniques ou le point de vue éthique évoluent, reconsidérer de façon éclairée l'usage que l'on peut faire d'un tel jeu de données.

Mais, on l'a vu, l'état actuel de la technologie ne permet pas de réutiliser une donnée pour le monde scientifique dès lors qu'elle possède un caractère personnel. Peut-être faut-il inventer un droit spécifique qui ferait de l'objet corpus un ensemble détaché des traces individualisables, de pouvoir le considérer comme l'on considère les résultats des cohortes en recherche médicale. Mais cela demanderait de résoudre un autre paradoxe : comment faire qu'un ensemble porte moins d'informations que la somme de ses parties ?

## Remerciements

Le projet ODISAE est financé par BPIfrance et la Région Ile de France, dans cadre du 17e Fonds Unifié Interministériel (FUI). Le projet réunit l'Aproged, La Cantoche Productions, le Centre Départemental du Tourisme de l'Aube, Jamespot, Kwaga, Eptica, l'INSEE, le Laboratoire d'Informatique de Nantes-Atlantique (LINA) et TokyWoky.

## Références

AMBLARD M. & FORT K. (2014). Étude quantitative des disfluences dans le discours de schizophrènes : automatiser pour limiter les biais. In *TALN - Traitement Automatique des Langues Naturelles*, p. 292–303, Marseille, France.

AMBLARD M., FORT K., MUSIOL M. & REBUSCHI M. (2014). L'impossibilité de l'anonymat dans le cadre de l'analyse du discours. In *Journée ATALA éthique et TAL*, Paris, France.

BOYD D. & CRAWFORD K. (2011). Six provocations for big data. In *A Decade in Internet Time : Symposium on the Dynamics of the Internet and Society*.

COUILLAULT A. & FORT K. (2013). Charte Éthique et Big Data : parce que mon corpus le vaut bien ! In *Linguistique, Langues et Parole : Statuts, Usages et Mésusages*, Strasbourg, France. 4 pages.

DE MAZANCOURT H., COUILLAULT A. & RECOURCÉ G. (2014). L'anonymisation, pierre d'achoppement pour le traitement automatique des courriels. In *Journée d'Etude ATALA Ethique et TAL*, Paris, France.

DE MONTJOYE Y.-A., HIDALGO C., VERLEYSEN M. & BLONDEL V. (2013). Unique in the crowd : The privacy bounds of human mobility. *Sci. Rep.*, **3**.

ESHKOL-TARAVELLA I., KANAAN-CAILLOL L., BAUDE O., DUGUA C. & MAUREL D. (2014). Procédure d'anonymisation et traitement automatique : l'expérience d'ESLO. In *Journée ATALA éthique et TAL*, Paris, France.

EUROPEAN COMMISSION (2014). The EU framework programme for research and innovation : How to complete your ethics self-assessment.

FORT K., ADDA G., SAGOT B., MARIANI J. & COUILLAULT A. (2014). Crowdsourcing for Language Resource Development : Criticisms About Amazon Mechanical Turk Overpowering Use. In Z. VETULANI & J. MARIANI, Eds., *Human Language Technology Challenges for Computer Science and Linguistics*, p. 303–314. Springer International Publishing.

KLIMT B. & YANG Y. (2004). Introducing the Enron corpus. In *First Conference on Email and Anti-Spam (CEAS)*.

MEDLOCK B. (2006). An introduction to nlp-based textual anonymisation. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06) : European Language Resources Association (ELRA).

# European Perspective on Privacy Issues in 'Free' Online Machine Translation Services.

Paweł Kamocki[1, 2, 3]
(1) Institut für Deutsche Sprache, R 5, 6-13, D-68161 Mannheim, Germany
(2) Institut Droit et Santé, Université Paris Descartes, 45 rue des Saints-Pères, 75270 Paris Cedex 6, France
(3) Institut für Informations-, Telekommunikations- und Medienrecht, WWU Münster, Leonardo-Campus 9, 48149 Münster, Germany kamocki@ids-mannheim.de

## Résumé.

**La perspective européenne sur les questions liées à la protection de la vie privée dans les outils 'gratuits' de traduction automatique en ligne**

Suite à la Révolution Numérique, la langue anglaise s'est établie comme la langue internationale; cependant, seuls 28,6% des internautes ont l'anglais pour langue maternelle. La traduction automatique (Machine Traslation, MT) est une technologie puissante qui peut combler ce fossé. En développement depuis le milieu du vingtième siècle, la traduction automatique est devenue accessible à chaque internaute au cours de la dernière décennie, grace aux outils disponibles gratuitement en ligne. Cette étude a pour ambition d'examiner les implications que ces outils peuvent avoir sur la vie privée des utilisateurs dans le contexte de la loi européenne sur la protection des données personnelles. Sont analysés le traitement initial (du point de vue de l'utilisateur et du fournisseur du service de traduction automatique) et le traitement secondaire qui peut potentiellement être entrepris par le fournisseur du service de traduction automatique.

**Abstract.** The English language has taken advantage of the Digital Revolution to establish itself as the global language; however, only 28.6% of Internet users speak English as their native language. Machine Translation (MT) is a powerful technology that can bridge this gap. In development since the mid-20th century, MT has become available to every Internet user in the last decade, due to free online MT services. This paper aims to discuss the implications that these tools may have for the privacy of their users and how they are addressed by EU data protection law. It examines the data-flows in respect of the initial processing (both from the perspective of the user and the MT service provider) and potential further processing that may be undertaken by the MT service provider.

**Mots-clés :** données personnelles, traduction automatisée, vie privée, Directive 95/46/CE, Google Translate
**Keywords:** personal data, Machine Translation, privacy, Directive 95/46/EC, Google Translate

## 1 Introduction

Digital revolution (started with the proliferation of personal computers in the late 1970s and continuing to the present day), just as any revolution worthy of its name, has changed our everyday life in more ways that one may want to admit. The switch from the analog to the digital has multiplied useful technologies that enabled an ordinary person to perform tasks that two generations ago required a considerable amount of time and/or manpower. Most importantly, new modes of communication developed in this Digital Age allow people to exchange information across the globe within seconds. A live chat with a contractor from another continent or an online search for the most obscure items sold in the four quarters of the Earth is now as easy as pie. Or is it...?

Not yet, and for a reason as old as hills: the language barrier. Even though English has undoubtedly taken advantage of the Digital Revolution to establish itself as a global language (Crystal, 1997); it has recently been estimated that it is used by 55,7% of all websites (German comes second with 6,1%)[1], as a matter of fact only 28,6% of Internet users speak English as a native language (as of December 31, 2013)[2]. While it is true that a certain percentage of the remaining Internet users speak (some) English as their second or third language, it remains a fact that a substantial part of the global Internet community does not speak it at all and, as a consequence, can only take advantage of a tiny fraction of the content available on the World Wide Web. Therefore, it cannot be denied that the global communication requires linguistic support systems in order to develop its full potential (Cribb, 2000).

[1] "Usage of content languages for websites," W3Techs, accessed April 13, 2015, http://w3techs.com/technologies/overview/content_language/all.
[2] "Internet World Users by Language," Internet World Stats, accessed April 13, 2015, http://www.internetworldstats.com/stats7.htm.

The Digital Revolution has provided a number of tools for linguistic support: it has (arguably) made language acquisition faster and more efficient, it has also helped improve the quality and the availability of human translation[3]. But none of these paths can lead to the erosion of language barriers in digital communication in the way that Machine Translation (MT) can accomplish this task.

## 1.1 MT in Context

MT (or automated translation) can be defined as a process in which software is used to translate text (or speech) from one natural language to another. This section will briefly present the history of MT and various technologies used in the process.

### *1.1.1 History*

The idea to mechanize the translation process can be traced back to the seventeenth century (Hutchins, 1986); most notably, in 1629 Descartes described a system of codes that would relate words between different languages (Hutchins, 1986), therefore allowing quick translation from one language to another. The first proposals for 'translating machines', however, did not appear until 1933, when Georges Artsrouni (Corbé, 1960) and Petr Troyanskii (Bel'skaya, Korolev Panov, 1959) were issued patents (in France and Russia respectively) for 'automated dictionaries'; both inventions remained nearly unknown and had become outdated before they were brought to the attention of the scientific community in 1950s-1960s.

In March 1949, inspired by the developments in code breaking during the Second World War, Warren Weaver, a researcher at the Rockeffeler Foundation, published a memorandum in which he put forward the idea to use computers for translation (Hutchins, 1999). This document marks the beginning of MT as a scientific discipline.

During the Cold War, researchers concentrated their efforts on Russian-to-English (in the US) and English-to-Russian MT (in the USSR). In January 1954 the first public demonstration of an MT system (used to translate more than sixty sentences from Russian to English) took place in the headquarters of IBM (the so-called Georgetown-IBM experiment — Hutchins, 2004). In the following years, imperfect MT systems were developed by American universities under the auspices of such players as the U.S. Air Force, Euratom or the U.S. Atomic Energy Commission (Hutchins, 1986).

At the end of 1950s, Yehoshua Bar-Hillel (the world's first full-time researcher in MT) questioned the possibility of developing a high-quality MT system, basing his argument mostly on semantic ambiguity of certain expressions in natural languages - a phenomenon that machines would never be able to deal with properly (Hutchins, 1998).

In 1964 the U.S. government, concerned about the lack of progress in the field of MT despite significant expenditure, commissioned a report from the Automatic Language Processing Advisory Committee (ALPAC). The report (the so-called ALPAC report), published in 1966, concluded that MT had no prospects of achieving the quality of human translation in foreseeable future (Hutchins, 1986). As a result, MT research was nearly abandoned for over a decade in the U.S.; despite these difficulties, the SYSTRAN company was established successfully in 1968 - their MT system was adopted by the U.S. Air Force in 1970 and by the Commission of the European Communities in 1976 (Hutchins, 1986).

During the 1980s, Japan found itself in the avant-garde of MT technology, with a particular focus on Japanese-to-English and English-to-Japanese translation (Hutchins, 1986).

In the 1990s, researchers (particularly in Germany) started to work on speech translation; during this period, low-end MT tools started appearing on personal computers. Most notably, in 1995 SYSTRAN launched SYSTRAN Professional for Windows. It was estimated that approximately 1000 MT packages were available for PCs in 1996 (Hutchins (ed.), 2000). BabelFish, the first online translation service (also based on SYSTRAN) was launched in 1997[4]. Google started providing an online translation (initially also using SYSTRAN) service in 2006[5] and Microsoft launched Bing Translator in 2009[6].

When American giants Microsoft and Google developed their proprietary MT systems in the second half of the 2000s (Google switched from SYSTRAN to a proprietary system in 2007; Microsoft Research developed an MT system in 2008 (Microsoft Translator Team, 2008), Europe also wanted to catch up. A German project Verbmobil (http://verbmobil.dfki.de/) was focused on German-English and German-Japanese language pairs. The Quaero project, established initially as a French-German cooperation, aimed at developing a multilingual search engine; MT technology was meant to be an important part of this initiative. After German partners announced their departure from the project, the future of this endeavour remains uncertain.

[3] a form of translation in which human translator uses software to facilitate the process is referred to as Computer-Assisted Translation (CAT) and should be clearly distinguished from Machine Translation (MT).
[4] according to the history of SYSTRAN software posted on www.translationsoftware4you.com.
[5] according to the company's history posted on Google's website.
[6] according to Wikipedia: http://en.wikipedia.org/wiki/Bing_Translator.

### *1.1.2 Technology and challanges*

The technological approach to MT has changed significantly over time. Chronologically the first MT method can be referred to as 'the dictionary method'. In this approach, words in a sentence are translated one-by-one, just like Descartes would have imagined it. This method, not different from a simple re-coding, can rarely produce satisfying results (although it is possible for closely related language pairs such as Dutch and English — Madsen, 2009); the system used in the IBM-Gorgetown experiment described above was in fact not much more complicated than this (the output translation was satisfying mostly because the input sentences were carefully selected — Madsen, 2009).

Everyone who has ever tried to translate a text knows that before it can be translated, it has to be understood first. The rules that humans use to decipher the meaning of linguistic expressions are morphological (i.e. how to build words out of morphemes) syntactic (i.e. how to build sentences out of words) or semantic (i.e. what words mean and what are the relations between them, such as e.g. hyponymy and hypernymy). According to some linguists, computers - in order to be able to provide MT of satisfying quality - have to integrate these rules (Winograd, 1972). This classic approach to MT is known as 'rule-based' or 'knowledge-based' MT.

A rule-based MT system, rather than translating word-by-word, performs a linguistic analysis of the input text (based on information retrieved from language-specific dictionaries, semantic hierarchies etc.) and then on the basis of its output generates a sentence in the target language.

Such a system is in fact not a simple tool, but a conglomerate of tools performing different tasks in this multi-stage process: a morphological analyser (in both source and target language), a syntactic parser, a thesaurus etc. Therefore, it is obvious that the development of such systems is costly and time-consuming, and that they are difficult and expensive to update. Nevertheless, such systems (the best known of which is SYSTRAN, especially in its older versions) were developed and implemented worldwide, starting from the 1970s.

With the development of corpus linguistics and the growth of computational power, another approach, the so-called 'statistical MT', has become more appealing. In this paradigm, revived in 1993 by IBM researchers (Brown *et al*., 1993), an MT system is based on a set of human-translated bilingual (or multilingual) language corpora. A statistical model derived from the analysis of the bilingual corpus is then used to translate input from source to target language according to the probability distribution.

The prerequisite for building statistical MT systems is therefore the existence of human-translated bilingual (or multilingual) corpora - and the bigger the better. Such corpora of satisfying quality may not exist for every language pair. An obvious source of human-translated multilingual corpora are international organizations such as the United Nations or the European Union, generating a substantial amount of freely available, high-quality multilingual documents (in 24 languages for the EU[7] and in 6 languages for the UN[8]).

Compared to rule-based MT systems, statistical MT systems are cheaper (at least for widely-spoken languages) and more flexible (a statistical system is not designed specifically for one language pair, but can accommodate to any language pair providing that the appropriate human-translated bilingual corpus can be 'loaded' into the system). Also, because statistical MT systems are based on human-translated texts, the output of statistical MT is (or at least can be) more natural. Finally, it can be assumed that, with the growth of available multilingual data, the quality of statistical MT will quickly improve.

In the last decade, there was a tendency for online MT providers to switch from the rule-based approach to the statistical approach (marked especially by Google Translate's shift from SYSTRAN to its proprietary, largely statistical MT system in 2007). Nowadays, the two approaches are often combined and it can be assumed that hybrid MT systems are the future of MT.

MT still has to face some important challenges, the most serious of which has always been disambiguation, i.e. choosing the 'right' meaning of an ambiguous input sentence. The two approaches described above can help the system to make the right choice, but they are still imperfect. It has also been put forward that MT cannot properly handle non-standard language, such as e.g. poetic expressions (but see: Genzel, Uszkoreit, Och, 2010); the expectations towards MT should therefore be mitigated. Some researchers claim even today that MT will never achieve the quality of human translation (Madsen, 2009).

Finally, the quality of MT output depends on the quality of the input. Even the most banal imperfections such as misspellings or grammar mistakes - not uncommon in electronic communications - even if they are barely noticeable to a human translator, can compromise the most elaborate MT system.

## 1.2 'Free' Online MT Tools

A number of 'free' online MT services are available today. This section will present the most popular of them and try to very briefly evaluate their quality.

---

[7] Art. 1 of the Regulation No. 1 of 15 April 1958 determining the languages to be used by the European Economic Community.

[8] Rule 51 of the Rules of Procedure of the General Assembly of the United Nations; rule 41 of the Provisional Rules of Procedure of the United Nations Security Council.

### *1.2.1 Examples*

The most popular 'free' online MT service is Google Translate. Launched in 2006, it can now support an impressive number of 80 languages, from Afrikaans to Zulu, including artificial (Esperanto) and extinct (Latin) languages. Google's proprietary MT system is based on the statistical approach. Google Translate is also available as an application for Android and iOS; it is integrated in Google Chrome and can be added as a plug-in to Mozilla Firefox.

Google Translate is the most popular 'free' online MT service; it's been reported to be used by 200 million people every day (in 2013[9]) and to to translate enough text to fill 1 million books every day (in 2012[10]).

SYSTRAN-based Yahoo! Babel Fish (formerly www.babelfish.yahoo.com) was chronologically the first 'free' online MT service. Opened by Systran and AltaVista in 1997, the service could support up to 38 languages. Alas, it no longer exists - in May 2012 it was replaced by Bing Translator. A company funded in 1995 named BabelFish still provides a 'free' online MT service at www.babelfish.com, supporting a humble number of 14 languages.

Bing Translator (http://www.bing.com/translator/) has been provided by Microsoft since 2009. It currently supports 44 languages and is integrated in Internet Explorer, Microsoft Office and Facebook.

An Israeli public company Babylon Ltd. and a French company Reverso-Softissimo also provides 'free' online MT tools. Babylon (http://translation.babylon.com) supports 30, and Reverso (www.reverso.net) 13 languages.

As a final example, a free/open-source MT platform Apertium is available at www.apertium.org. The engine, the tools and the data are licensed under CC BY-SA 3.0 or GNU GPL 3.0 and can be freely shared and re-used. Unfortunately Apertium offers mostly translation for closely-related language pairs.

### *1.2.2 Are they '**good enough**' ?*

Erik Ketzan argued in 2007 that the fact that MT had not attracted much attention from legal scholars was a consequence of the low quality of the output (Ketzan, 2007). He predicted that '*if MT ever evolv[ed] to "good enough," it [would] create massive copyright infringement on an unprecedented global scale'*. While this article is not about copyright issues in MT, the question 'is MT good enough?' remains relevant.

First of all, given that organizations such as U.S. Air Force or the European Commission use high-end MT systems we assume that the technology is (and was already in the 1970s) far from being useless altogether. Whether low-end, 'freely' available online MT tools are 'good enough' is a slightly different issue.

The user's expectations related to such services have to be reasonable, but we believe that they can be satisfied to a large extent. 'Free' online MT tools can definitely help users understand e-mails or websites in foreign languages. Moreover, we remark that the quality of MT tools is improving. For example, in his article Ketzan quoted '*My house is its house'* as machine translation of the Spanish proverb '*Mi casa es su casa'*. He obtained this result on August 6, 2006 using Google Translate (which was at that time based on SYSTRAN). Today (April 15, 2015) the proverb is translated correctly as '*My house is your house'*.

If we take into account Moore's law (according to which computers' speed and capacity double every 18 months — More, 1965)[11], as well as the exponentially growing number of digital language data that may be used to increase the accuracy of statistical MT systems, the future of MT technology looks promising. The number of users of 'free' online MT services will probably keep growing -- it is therefore important to discuss the impact that such tools may have on their privacy.

## 2 Nature of the Data Processed in 'Free' Online MT Tools

'Free' online MT services can be used to process all sorts of texts - private and professional correspondence, commercial offers, blogs, Internet fora, social media content. The input may be of different length -- from several paragraphs (e.g. Bing Translator is limited to 5000 characters, but Google Translate can handle several times more) to single words.

In the light of the above it is not surprising that information entered by users in 'free' online MT tools can be sensitive from the point of view of privacy. In fact probably much more often than the users would like to acknowledge it, the input may be qualified as personal data.

The concept of personal data is defined in art. 2(a) of the Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data (hereinafter: the Directive). For the purpose of the Directive, personal data shall mean '*any information relating to an identified or identifiable natural person'*. According to Article 29 Data Protection Working Party (hereinafter: WP29), the definition should be interpreted quite broadly. For example, it covers not only the 'objective' information, but also 'subjective' information such as opinions or assessments[12]. Moreover, the information 'relates to a person' not only if it is 'about' a person (the 'content' element), but also if it is used to evaluate or influence the status or behavior of the person (the 'purpose'

[9] according to http://www.cnet.com/news/google-translate-now-serves-200-million-people-daily/.
[10] according to http://googleblog.blogspot.co.uk/2012/04/breaking-down-language-barriersix-years.html.
[11] this growth rate, however, which continued steadily for more than half a century, is expected to stop very soon.

element), or if it has an impact on the person's interests or rights (the 'result' element). The person that the data relate to (i.e. the data subject) can be not only identified, but also identifiable by any means likely reasonably to be used by the data controller or any other person[13]. It has to be taken into account here that the providers of 'free' online MT tools often provide a wide range of internet services (email, social media platforms, applications...); they may therefore be in possession of vast datasets which, cross-referenced with the data entered into an MT system, may help identify the data subject[14]. Finally, the data processed in 'free' online MT tools may concern not only natural, but also legal persons (e.g. information on the financial condition of an employer can be found in the employee's private e-mails), which would exceed the definition of personal data, but may still be concerned by privacy/confidentiality considerations.

Users might be inclined to believe that what MT tools do is nothing more than simple automatic re-coding; in fact, as it has been shown above, MT tools perform an analysis of the input text. This analysis can certainly be qualified as 'processing' in the sense of art. 2(b) of the Directive[15].

In our view, the processing of data in 'free' online MT tools can be divided into two stages: first, the information is entered in order to be translated (this stage can be called 'initial processing'); then, the MT provider may want to further process the input data for different purposes (these purposes may range from scientific research on the development of the system, the evaluation of the system, statistics to direct marketing). The following sections will examine these two stages in the context of the Directive.

# 3 Initial Processing

The two actors at this stage of processing are: the user (who enters the data into the system) and the MT provider. It is not clear whether both of them can be regarded as data controllers in the sense of the Directive[16], or whether only the user is the controller, and the provider - a processor[17]. In our view - given that the provider plays a crucial role in determining the functioning of his MT system - both actors can be regarded as processors[18].

## 3.1 The User's Perspective

By entering data into the MT system the user may not only process information concerning him, but also information concerning other natural (or even legal) persons. It seems, however, that merely entering the data into a freely available MT tool for the purpose of obtaining an imperfect translation (to be able to better understand the text) may in most cases be qualified as a 'purely personal or household activity' and as such exempted from the Directive under art. 3(2).

It is not clear how to understand 'purely personal or household activities'. In our view, the use of freely available MT tools may fall in the first category even if the tool is used for professional or academic purposes (in fact, the Directive mentions 'private activity', and not 'private purposes'), unless the user is a professional translator. It is difficult to argue that human-translating a text without publishing the translation - just to be able to understand the original - may amount to unfair personal data processing, even if the text contains someone else's personal information. This should also be the case if MT is used.

## 3.2 The Provider's Perspective

By definition, machine translation is carried out in an automated way. Of course, the Directive still applies to such processing (art. 3(1)).

International MT providers such as Google or Microsoft could argue that the Directive does not apply to them because they are not established on the territory of an EU Member State nor do they use equipment situated on the territory of such a state (art. 4 of the Directive). This argument (which has already been rejected by European courts and data protection authorities)[19], however, will soon become invalid as the proposed text of the new General Data Protection

[12] Article 29 Data Protection Working Party, Opinion 4/2007 on the concept of personal data adopted on 20th June 2007, 01248/07/EN, WP 136, 6.

[13] Recital 26 of the Directive 95/46/EC.

[14] see: the concept of 'linkability' in: Article 29 Data Protection Working Party, Opinion 05/2014 on anonymisation techniques adopted on 10 April 2014, 0829/14/EN, WP 216, 11.

[15] 'processing' is defined in art. 2(b) of the Directive 95/46/EC as *'any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction'*.

[16] Art. 2(d) of the Directive 95/46/EC defines the controller as *'the person who determines (alone or jointly with others) the purposes and means of the processing of personal data'*.

[17] Art. 2(e) of the Directive 95/46/EC defines the processor as *'the person which processes personal data on behalf of the data controller'*.

[18] cf.: CJEU judgement in Case C-131/12 Google Spain SL, Google Inc. v Agencia Espan#ola de Protección de Datos, Mario Costeja Gonzàlez, par. 23: *'The operator of a search engine is the 'controller' in respect of the data processing carried out by it since it is the operator that determines the purposes and means of that processing'*.

Regulation extends its applicability to the processing activities related to the offering of services (such as MT services) to data subject in the EU (Tene, Wolf, 2013).

Finally, MT providers may try to invoke a liability limitation, e.g. under art. 14 of the Directive 2000/31/EC on e-commerce[20], claiming that all they do is to provide a service that consists of processing data provided by the users. This argument also has little chance of success in court -- the CJEU held recently[21] that search engine providers are responsible for the processing of personal data which appear on web pages published by third parties.

Therefore, MT providers are bound by the provisions of the Directive, such as those according to which processing may only be carried out on the basis of one of the possible grounds listed in its art. 7. In our view, the only two grounds that can be taken into consideration here are: the data subject's consent (art. 7(a)) and performance of a contract to which the data subject is party (art. 7(b)).

#### *3.2.1 Consent*

Art. 2(h) of the Directive defines 'consent' as '*any freely given specific and informed indication of his wishes by which the data subject signifies his agreement to personal data relating to him being processed'*. According to WP29, it can be '*any signal, sufficiently clear to be capable of indicating the data subject' s wishes and to be understandable by the data controller'*[22]. It seems therefore that consent may be concluded from the data subject's behavior. The use of an MT service can, in our view, be interpreted as consent for processing, but only to the extent necessary for the translation (and not for any further processing).

#### *3.2.2 Performance of a contract*

Another ground for lawfulness that can be imagined in the context of 'free' online MT services is performance of a contract to which the data subject is party.

The MT provider offers an MT service to the users; by entering data in the service in order to obtain a translation the user accepts the offer. Without entering into details of contract law theory, we believe that these circumstances may be sufficient for a contract to be formed, at least in jurisdictions that do not require consideration (i.e. something of value promised to another party) as a necessary element of a contract (however, the data itself may be regarded as consideration for the translation service -- see below).

The processing of data is therefore necessary for the performance of such a contract - which in itself may constitute a valid ground for lawfulness.

## 4 Further processing

Contrary to what some users may imagine, the data entered in a 'free' online MT system do not 'disappear' once the MT task is accomplished. In fact, some MT providers expressly state in their Terms of Service that by entering data in their services the users grant them a copyright license[23], which indicates the provider's intention to re-use the data.

In our view, the purposes for which MT providers may further process the data may be divided into two categories. Firstly: non-commercial purposes (including statistics, evaluation and improvement of the system) and secondly: commercial purposes (including e.g. direct marketing).

### 4.1 Non-commercial Purposes

This section examines possible grounds for further processing of data entered into online MT services for non-commercial purposes. In our view, these grounds include: processing for purposes compatible with the initial consent and legitimate interest of the data controller. In addition, such processing may also be carried out on the basis of a statutory exception.

#### *4.1.1 Purpose Limitation*

According to art. 6(1)(b) of the Directive, personal data must be 'collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes'. This principle, referred to as 'purpose

[19] cf.: CJEU judgement in Case C-131/12; Déliberation de la Commission Nationale Informatique et Libertés no. 2013-420 prononçant une sanction pécuniaire à l'encontre de la société Google Inc.

[20] this provision has received rather extensive interpretation from the CJEU, especially in case Google France SARL and Google Inc. v Louis Vuitton Malletier SA (C-236/08) concerning the AdWords service.

[21] in case C-131/12 (cf. *supra*).

[22] Article 29 Data Protection Working Party, Opinion 15/2011 on the definition of consent adopted on 13 July 2011, 01197/11/EN, WP187, 11.

[23] cf. e.g. Google's Terms of Service, last modified April 14, 2014, http://www.google.com/intl/en/policies/terms/.

limitation', allows further processing for purposes compatible with the initial purpose (provided that this purpose is specified, explicit and legitimate).

As discussed above, MT can be analyzed as a specified (i.e. sufficiently defined), explicit (i.e. unambiguous) and legitimate (i.e., among others, carried out on the basis of one of the grounds listed in art. 7) purpose[24]; the possible grounds for legitimacy are consent of the data subject and performance of a contract to which the data subject is party (see above).

The next step is to assess whether the purposes such as evaluation and development of the tool or scientific research can be regarded as compatible with the initial purpose. According to WP29[25], the key factors to be considered during the compatibility assessment include:

*– the relationship between the initial purpose and further purposes:*

In our view, evaluation and improvement of the MT tool (further purposes) are very closely related to the translation (initial purpose) - indeed, it can be said that such further processing may be regarded as a logical next step of the initial processing.

*– the context in which the data have been collected and the reasonable expectations of the data subjects as to their further use:*

We assume that an average reasonable person imagines that the data entered in a 'free' online MT system 'disappear' once the task is accomplished (Porsiel, 2012) and does not expect them to be further processed.

A link to a document containing detailed information about the service's privacy policy does not seem to influence the assessment of this factor: it is no secret that an average user does not read such documents.

It has to be noted, however, that the awareness and reasonable expectations of users may vary over time. While nowadays an average user does not seem to realize that his personal data are in fact a currency that he can use to pay for 'free' online services (European Data Protection Supervisor, 2014), it may become obvious a decade from now. This may be the case especially if, being aware of the existence of payable alternatives, the user chooses deliberately to use a 'free' service[26].

*– the nature of the data and the impact of the further processing on the data subjects:*

This factor is difficult to assess a priori, as it depends on the circumstances of each case. In fact, circumstances in which a 'free' online MT service is used to translate texts containing sensitive information (e.g. about health, sex life, religious or philosophical beliefs etc.)[27] cannot be excluded. Whether the processing of such data for non-commercial purposes mentioned above can have adverse consequences for the data subject is another issue.

According to WP29, in assessing the possible impact on the processing on the data subject, several elements different from the nature of the data should also be taken into account. These include: whether the data are processed by a different controller, publicly disclosed to a large number of persons or combined with other data. In our view, further processing strictly limited to the 'non-commercial' purposes mentioned above may pass this test; this will definitely not be the case of processing for commercial purposes.

*– the safeguards applied by the controller to ensure fair processing and to prevent any undue impact on the data subject:*

Once again, the assessment on this factor will be different for different MT systems. While some providers of paid MT services respect high security standards (such as ISO/IEC 27001:2013), this may not be the case of all 'free' MT services.

The experience shows that even companies believed to conform to high security standards may suffer from 'privacy incidents'; their approach to 'privacy by design' is not flawless (Rubinstein, Good, 2013), and their policies and practice in this field have been criticized for lack of compatibility with EU law[28].

An important element to take into consideration here is pseudonymization or anonymization of the data. The extent to which MT service providers apply these safeguards remains uncertain (Toubiana, Nissenbaum, 2011).

As far as processing for research purposes is concerned, art. 6(1)b of the Directive contains a specific provision on further processing for 'historical, statistical or scientific purposes'. It enables the Member States to authorize such processing, as long as appropriate safeguards are provided. In practice, it seems that Member States set the threshold for

[24] Article 29 Data Protection Working Party, Opinion 03/2013 on purpose limitation adopted on 2 April 2013, 00569/13/EN, WP 203, 12.

[25] *idem*

[26] Article 29 Data Protection Working Party, Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC adopted on 9 April 2014, 844/14/EN, WP 217, 47.

[27] Art. 8(1) of the Directive 95/46/EC.

[28] cf.: Déliberation de la Commission Nationale Informatique et Libertés no. 2013-420; WP29's letter to Larry Page (CEO of Google Inc.) of 23 September 2014, accessed October 23, 2014, http://ec.europa.eu/justice/data-protection/article-29/documentation/other-document/files/2014/20140923_letter_on_google_privacy_policy.pdf.

'appropriate safeguards' rather high[29]; therefore, it seems that providers of 'free' online MT services will rarely be able to meet it.

In the light of the above, it remains uncertain whether further processing of data for non-commercial purposes can be regarded as compatible with the initial purpose. In fact, it may be easier for MT providers to rely on art. 7(f) to legitimize such processing.

#### *4.1.2 Legitimate Interests*

Art. 7(f) of the Directive provides an open-ended ground for privacy by allowing processing of personal data '*necessary for the purposes of the legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed*'.

WP29 provides guidelines for carrying out the balancing test: in order to be regarded as 'legitimate', the controller's interest has to be lawful, sufficiently concrete and it has to represent real and present interest[30]. In our view, the evaluation and the development of the tool which can be achieved via additional processing of input data can pass this test.

Moreover, the processing has to be necessary to achieve the interest pursued. The word 'necessary' shall not be interpreted as synonymous with 'indispensable'[31]; instead, it should be examined whether there are other less invasive means to achieve the intended purpose[32]. It may seem at first sight that indeed there are other ways of improving or evaluating an MT system than by enlarging the corpus that the system is based on. Also, the data necessary for the enlargement of the corpus may well be obtained from different sources. This part of the test may therefore be difficult to pass.

Establishing balance between the interests of MT providers and the rights of the data subject is not an easy task; many factors, some of them similar to those examined above in the context of the purpose limitation principle, should be taken into account. In our opinion, the fact that the benefits that the whole community may derive from the advancement in free online MT are so important that they outweigh the interests of the data subject - especially in view of the fact that the risks of adverse consequences for the data subject remain relatively low.

The fact that it is definitely not easy - if possible - for data subjects to exercise their right to object to the processing does not work in favor of MT service providers. The recent judgement of the CJEU[33], however, allows us to believe that the right to be forgotten will soon have to become a standard for online services, including MT.

To conclude, it is not clear whether MT providers can further process personal data entered into their MT systems for the purposes of research, improvement and evaluation of these systems. It is our belief, however, that the interest of the whole Internet community in the development of free online MT tools outweighs the possibly negative impact that such processing may have on the data subject. Therefore, in our view, art. 7(f) of the Directive may be a sufficient ground for lawfulness of such processing. Of course, other principles relating to data processing (laid down in art. 6 of the Directive) should still be respected.

#### *4.1.3 Research exception*

Unlike the Directive, the new General Data Protection Regulation contains (in its art. 83) a research exception[34]. According to the wording proposed by LIBE in October 2013, such processing can be allowed if it passes the necessity test (i.e. if its purposes cannot be achieved by processing anonymized data - see above), or if the information allowing identification of the data subject is separated from the data and kept separately under the highest technical standards. It is too early to say how this provision can be interpreted in practice (e.g. whether compliance with an ISO standard will be necessary to meet the 'highest technical standards' requirement). In any case, it seems that it will not bring much relief to MT providers, unless they implement robust anonymization (or at least pseudonymization) mechanisms.

### 4.2 Commercial purposes

MT providers may want to further process the data entered into MT services for purposes such as direct marketing (e.g. sending advertisements of local Italian restaurants for those who had used the free online MT service to translate a recipe from Italian), consumer profiling etc. All these purposes will be jointly referred here as 'commercial'.

---

29 cf. e.g. in Germany: art. 13(2)8 BDSG (requiring that the scientific interests must significantly outweigh the interests of the data subjects), in France: chapter IX CNIL (containing a derogatory framework for health research).

30 Opinion 06/2014 (cf. *supra*), 25.

31 Judgement of the European Court of Human Rights in Silver & Others v. United Kingdom of 25 March 1983.

32 Opinion 06/2014 (cf. *supra*), 29.

33 in case C-131/12 (cf. *supra*).

34 Proposal for a regulation of the European Parliament and of the Council on the protection of individual with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Compromise Amendments on art. 30-91: COMP Art. 83 of 17.10.2013.

According to the rules presented in the preceding section, it is obvious that the data collected from users of MT tools cannot be further processed for such purposes.

In particular, the argument that by using an MT service the user gives his consent for such processing just because the Terms of Use or the Privacy Policy stipulate so must be declared invalid[35]. Such consent could not be regarded as sufficiently informed and would not meet the requirements of the Directive.

Other possible grounds (such as art. 7(f) of the Directive - see above) would also have to be rejected. Finally, commercial purposes cannot be regarded as compatible with the initial purpose of the processing (translation).

# 5 Conclusions

While 'free' online MT services provide convenient and quick translation of rather satisfying quality, they may pose a threat to privacy of users. In fact, data entered into such a service do not disappear once the translation is accomplished; instead, the MT providers may want to use the data for various purposes including not only the evaluation and improvement of the tool, but also direct marketing.

In our view, the existing EU data protection framework in itself is not an obstacle to the functioning of such services - the exemption of 'purely private activities' and the possibility to interpret certain behaviors of data subjects as consent for processing seem to strike balance between the interests of the user and those of the MT provider.

Furthermore, in our view the existing framework provides for sufficient protection against further processing of the data for 'commercial' purposes. Whether these rules are respected in practice, however, is a different issue.

# Acknowledgements

The author would like to thank Dr. Marc Stauch, Mr. Jim O'Regan and Dr. Piotr Bański for their expert advice and encouragement.

# References

BEL'SKAYA I. K., KOROLEV L. N., PANOV D. YU. (1959). *Переводная машина П. П. Троянского: сборник материалов о переводной машине для перевода с одного языка на другие, предложенной П. П. Троянским в 1933 г.*. Moscow: Изд. Акад. Наук.

BROWN P. F., DELLA PIETRA S. A., DELLA PIETRA V. J., MERCER R. L. (1993). The mathematics of statistical machine translation: parameter estimation, *Computational Linguistics* 19, 263-311.

CORBÉ M. (1960). La machine à traduire française aura bientôt trente ans. *Automatisme* 5, 87-91. CRIBB M. V. (2000). Machine Translation: The Alternative for the 21st Century? *TESOL Quaterly* 34, 560-569. CRYSTAL D. (1997). *English as a global language.* Cambridge: University Press.

EUROPEAN DATA PROTECTION SUPERVISOR (2014). *Privacy and competitiveness in the age of big data: The interplay between data protection, competition law and consumer protection in the Digital Economy*, Brussels : European Data Protection Supervisor.

GENZEL D., USZKOREIT J., OCH F. (2010). « Poetic » Statistical Machine Translation: Rhyme and Meter. Proceedings of the *2010 Conference on Empirical Methods in Natural Language Processing, MIT, Massachusetts, USA, 9-11 October 2010*, 158–166.

HUTCHINS J. W. (1986). *Machine translation: past, present, future*. New York: Halsted Press. HUTCHINS J. W. (1998). Bar-Hillel's survey, 1951. *Language Today* 8, 22-23.

HUTCHINS J. W. (1999). Warren Weaver memorandum: 50th anniversary of machine translation. *MT News International* 22, 5-6.

HUTCHINS J. W. (ed.) (2000). *Compendium of Translation Software, Machine Translation Systems and Computer-aided Translation Support Tools. 1st edition.* Geneva : European Association for Machine Translation.

[35] cf. Article 29 Data Protection Working Party, Opinion on the use of location data with a view to providing value-added services adopted on November 2005, 2130/05/EN, WP115, 5: *'This definition* [in art. 2(h) of the Directive] *explicitly rules out consent being given as part of accepting the general terms and conditions for the electronic communications service offered'* ; see also Déliberation de la Commission Nationale Informatique et Libertés no. 2013-420.

HUTCHINS J. W. (2004). The Georgetown-IBM experiment demonstrated in January 1954. Proceedings of the *6th Conference of the Association for Machine Translation in the Americas, AMTA 2004, Washington, DC*, 102-114.

KETZAN E. (2007). Rebuilding Babel: Copyright and the Future of Machine Translation Online. *Tulane Journal of Technology & Intellectual Property* 9, 205-234.

MADSEN M. W. (2009). *The Limits of Machine Translation*. PhD diss., University of Copenhagen. MOORE G. E. (1965). Cramming More Components onto Integrated Circuits. *Electronics* 38, 114-117.

RUBINSTEIN I., GOOD N. (2013). Privacy by Design: A Counterfactual Analysis of Google and Facebook Privacy Incidents. *Berkeley Technology Law Journal* 28, 1322-1414.

TENE O., WOLF CH. (2013). *Overextended: Jurisdiction and Applicable Law under the EU General Data Protection Regulation*. Washington : Future of Privacy Forum.

TOUBIANA V., NISSENBAUM H. (2011). Analysis of Google Logs Retention Policies. *Journal of Privacy and Confidentiality* 3, 3-26.

WINOGRAD T. (1972). *Understanding Natural Language.* New York: Academic Press.

# Annotateurs volontaires investis et éthique de l'annotation de lettres de suicidés

K. Bretonnel Cohen[1] John P. Pestian[2] Karën Fort[3]
(1) University of Colorado, Denver
(2) Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati
(3) Université Paris-Sorbonne / STIH
kevin.cohen@gmail.com, john.pestian@cchmc.org, karen.fort@paris-sorbonne.fr

**Résumé.** Cet article présente une perspective éthique sur le projet décrit dans (Pestian *et al.*, 2012b). La campagne d'annotation en question a visé à produire un corpus de lettres de suicidés annotées en émotions. Les annotateurs étaient soit des parents ou des amis de suicidés, soit des professionnels de la santé mentale. Nous appelons ces annotateurs bénévoles, volontaires pour faire avancer la recherche, des *volontaires investis*. Ce projet soulève un certain nombre de questions éthiques, notamment en ce qui concerne le rôle de l'empathie des annotateurs, les effets possibles sur ceux-ci et les utilisations potentielles des résultats obtenus. Nous concluons par une analyse du corpus du point de vue de la *Charte Éthique et Big Data*.

**Abstract.**
**Annotating suicide notes : ethical issues at a glance.**
According to the World Health Organization, 800,000 people die of suicide every year. About 20% of them leave a written message. This paper discusses a corpus of such messages. The corpus was annotated with reference to the emotions expressed in the notes. The annotators were family or friends of someone who had died by suicide, or mental health professionals. We refer to these non-coercivally and altruistically motivated annotators as *vested volunteers.* A number of ethical issues are explored with this task and group of annotators, including the role of empathy, possible effects on the annotators, and the uses that might be made of the products of the annotation project. We conclude considering the project from the point of view of the *Ethics and Big Data Charter*.

**Mots-clés :** lettres de suicidés, éthique, annotation, myriadisation, corpus.

**Keywords:** suicide notes, ethics, annotation, crowdsourcing, corpus.

## 1 Introduction

Selon l'Organisation Mondiale de la Santé (OMS), chaque année dans le monde, plus de 800 000 personnes se donnent la mort [1]. Environ 20 % d'entre elles laissent un message écrit (on peut citer par exemple, (Volant, 1990) 664 sur 3 450 et (Fédération française de psychiatrie, 2001), 200 sur 621). On parle en anglais de *suicide note*. Nous utiliserons ici le terme de *lettre*, reprenant à notre compte la remarque à ce sujet d'Éric Volant (Volant, 1990) :

> Elles sont bien davantage qu'un simple mot qu'on laisse traîner [...]. Même s'il ne s'agit que d'un griffonnage, il a un but défini pour son auteur, et mérite donc pleinement le statut de lettre.

Un certain nombre de chercheurs ont étudié différents aspects de ces lettres, dans le but d'aider à la prévention du suicide. L'analyse que nous présentons ici concerne un projet de recherche de ce type, mené aux États-Unis, qui a donné lieu à la constitution et à l'annotation en émotions d'un corpus de lettres de suicidés. Ce projet est détaillé dans (Pestian *et al.*, 2012b).

Les lettres de ce corpus ont été annotées par des volontaires qui ont eu une expérience concrète du suicide, soit parce qu'un membre de leur famille ou un de leurs amis s'est suicidé, soit parce qu'ils sont des professionnels de la santé mentale.

Ceux-ci ont participé au projet bénévolement, sans contrainte, pour faire avancer les travaux des scientifiques dans ce domaine. Pour rendre compte à la fois de cette volonté et de leur rapport personnel au suicide, nous les appelons « volontaires

1. Voir http://www.who.int/mediacentre/news/releases/2014/suicide-prevention-report/fr/.

investis » (*vested volunteers*).

Ce type de projet pose un certain nombre de questions éthiques, que l'appel à des volontaires investis rend encore plus saillantes. Il en va ainsi du rôle de l'empathie de l'annotateur, des effets potentiels sur celui-ci, des sources de biais et des utilisations possibles des résultats de l'étude.

## 1.1 Une foule (limitée) de volontaires

Le terme anglais *crowdsourcing* est défini par le dictionnaire en ligne Merriam-Webster comme « la pratique consistant à obtenir le service désiré, qu'il s'agisse d'une idée ou de contenu, en sollicitant les contributions d'un grand groupe de personnes, en particulier de la communauté des internautes plutôt que des employés ou fournisseurs traditionnels [2] ». Ce mot valise très expressif, formé de foule (*crowd*) et de délocalisation (*outsourcing*), se délave à la traduction. Nous le traduirons ici par myriadisation [3].

Ces dernières années, la myriadisation est devenue une méthode populaire de construction de ressources langagières, en particulier dans le cadre de la recherche en Traitement Automatique des Langues (TAL). Cet essor n'a pas été sans poser des questions éthiques. Ainsi, (Fort *et al.*, 2011) présente une critique détaillée de l'utilisation d'`Amazon Mechanical Turk`, une plate-forme de travail parcellisé qui permet l'exploitation de travailleurs de pays en développement. Les conditions de travail des *Turkers* ont d'ailleurs fait l'objet d'études sur le terrain qui confirment cette exploitation (Gupta *et al.*, 2014).

D'autres méthodes de myriadisation, plus satisfaisantes éthiquement, se sont développées en parallèle. Parmi celles-ci, les jeux ayant un but (*Games With A Purpose*) proposent de produire des ressources en jouant. Le premier de ces jeux pour les ressources langagières est à notre connaissance `JeuxDeMots` (Lafourcade & Joubert, 2008), qui crée un réseau lexical en français. `Phrase Detectives` (Chamberlain *et al.*, 2008) a quant à lui été utilisé pour annoter des anaphores dans un corpus en anglais. Plus récemment, `ZombiLingo` (Fort *et al.*, 2014) permet l'annotation en syntaxe de dépendances d'un corpus en français.

Le point commun entre `JeuxDeMots`, `ZombiLingo` et d'autres formes de myriadisation comme `Wikipédia` ou le `Projet Gutenberg` est le bénévolat. Les participants savent qu'ils ne seront pas rémunérés [4]. Ils sont motivés par l'aspect ludique du jeu ou par l'envie de participer à un projet de bien commun ou de recherche. Dans ce dernier cas, on parle alors de sciences participatives. Un exemple de projet de science participative sans aspect particulièrement ludique est `Vigie Nature` développé au Muséum National d'Histoire Naturelle (Couvet *et al.*, 2011).

À la différence de ces projets, la campagne d'annotation en discussion ici n'a rien de plaisant. La motivation des participants est donc simplement d'aider à faire avancer la recherche sur le sujet du suicide. Une autre différence est que l'appel à participation a été limité à certaines communautés (voir section 2.2) et n'a pas été totalement ouvert.

## 1.2 Une annotation discriminante

Les lettres de suicidés ont fait l'objet d'un nombre de recherches considérable (Shneidman & Farberow, 1957b; Osgood & Walker, 1959; Tuckman *et al.*, 1959; Schneidman, 1981; Leenaars, 1988; Baume *et al.*, 1997; Brevard *et al.*, 1990; Ho *et al.*, 1998; O'Connor & Leenaars, 2004; Olson, 2005).

Le domaine tel que nous le connaissons aujourd'hui a vu le jour lorsque Edwin Shneidman a découvert un gisement de lettres de suicidés dans les bureaux du Coroner du comté de Los Angeles. Sa découverte a été complétée par la création d'un corpus témoin : il a demandé (pour des raisons pratiques) à des travailleurs syndiqués d'écrire la lettre qu'ils écriraient s'ils voulaient se suicider (Shneidman & Farberow, 1957a). La notion de contrôle en suicidologie est complexe (Pestian, 2010), les participants « témoins » ont cependant été sélectionnés pour correspondre aux suicidés en termes d'âge, de genre, de religion et de nationalité (tous américains).

Ces données ont servi dans de nombreuses expériences qui ont montré qu'« il est possible de distinguer entre de vraies lettres de suicidés et des simulacres, et, encore plus important, que les lettres de suicidés se caractérisent généralement par une logique dichotomique, davantage d'hostilité et d'auto-critique, de noms spécifiques et d'instructions aux survivants,

2. *the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community rather than from traditional employees or suppliers* (Merriam-Webster, consulte en avril 2015)

3. Cette traduction a été proposée par Gilles Adda dans (Sagot *et al.*, 2011).

4. Le cas de `Phrase Detectives` est un peu différent, car les joueurs peuvent gagner des bons d'achat.

| | Moy. des lettres (écart type) | Moy. des lettres simulées (écart type) | Signif. |
|---|---|---|---|
| Nb phrases | 9,242 (8.058) | 4,848 (3,308) | 0,001 |
| Fréquence max. d'un mot | 7,909 (6,090) | 4,212 (3,525) | 0,002 |
| Nb caractères hors mot | 13,182 (11,078) | 7,212 (7,039) | 0,004 |
| Niveau de lisibilité Flesch-Kinkaid | 4,719 (2,142) | 6,517 (2,994) | 0,008 |

TABLE 1 – Quelques spécificités des lettres de suicidés par rapport à des lettres simulées.

moins de signes de réflexion sur ce que l'autre pense, et plus d'usage des différents sens du mot "amour" » (Shneidman, 1973).

Nous présentons ci-dessous deux lettres d'adieux au monde [5] écrites par des hommes américains :

Please call my children first Can't Take much more Love you all to much gone to See Mary I can't live without Mary this pain + misery is to much Love all you kids forgive me ain't going to Die in Hosp. Like Mom Love Dad

(Leenaars, 1988)

I'm sorry, but somewhere I lost the road, and in my struggle to find it again, I just got further and further away. There should be little sadness, and no searching for who is at fault; for the act and the result are not sad, and no one is at fault. My only sorrow is for my parents who will not easily be able to accept that this is so much better for me. Please folks, it's all right, really it is. 1 :30 p.m.-The ultimate adventure begins! Car to Helen or Ray (needs a tune up). Money to Max and Sylvia. Furniture to George. I wanted to be too many things, and greatness besides-it was a hopeless task. I never managed to learn to really love another person-only to make the sounds of it. I never could believe what my society taught me to believe, yet I could never manage to quite find the truth. 2 :15 p.m.-:-I am about to will myself to stop my heart beat and respiration. This is a very mystical experience. I have no fear. That surprises me. I thought I would be terrified. Soon I will know what death is like-how many people out there can say that?

(Shneidman, 1973)

La première de ces lettres est relativement typique, puisqu'elle contient des verbes à l'impératif et l'expression d'une souffrance.

La seconde a été écrite par un psychiatre de 31 ans qui s'est donné la mort en prenant des barbituriques. Il a été trouvé dans une forêt du mid-Ouest américain. Cette deuxième lettre est moins typique, car elle ne mentionne pas de souffrance (du moins pas explicitement). Elle contient cependant elle-aussi des formes à l'impératif.

Le tableau 1 présente un certain nombre d'autres caractéristiques des véritables lettres de suicidés par rapport à des lettres simulées. Nous nous limitons ici aux spécificités les plus significatives ($p < 0{,}01$ ou moins).

Il est à noter que les lettres prises en compte dans ces chiffres ont été écrites par des Américains et qu'il faut donc se garder de généraliser à d'autres cultures, y compris d'autres pays anglo-saxons.

Le corpus de lettres dont il est question ici a été annoté en émotions, afin d'affiner l'analyse de ces lettres et aller au-delà de la simple classification entre « vraie » et « fausse » lettre de suicidé.

# 2 Matériels et méthodes

## 2.1 Création du corpus

Les 1 319 lettres du corpus ont été collectées entre 1950 et 2012 par Edwin Shneidman (UCLA), et John Pestian, du Centre médical des enfants de l'hôpital de Cincinnati (CCHMC). La mise en place de la base de données a commencé

5. Nous utilisons ici à dessein le terme employé par (Volant, 1990).

en 2009 et a reçu l'aval des institutions (CCHMC's Institutional Review Board). Chaque lettre a été scannée dans le module spécialisé (*Suicide Note Module*) de la plate-forme de décision clinique CHRISTINE. Les lettres ont ensuite été transcrites par un transcripteur professionnel et chacune a été relue par trois relecteurs en parallèle. Leurs instructions étaient de corriger les erreurs de transcription mais de laisser les erreurs d'origine, telles que les fautes d'orthographe, de grammaire, etc.

Afin de préserver la vie privée des suicidés, les lettres ont ensuite été anonymisées. Afin qu'elles restent utilisables par des systèmes par apprentissage supervisé, les informations retirées ont été remplacées par des valeurs équivalentes qui protègent la vie privée des individus. Ainsi, tous les prénoms de femmes ont été remplacés par « Jane » et tous les prénoms d'hommes par « John ». Tous les noms de famille ont quant à eux été remplacés par « Johnson ». Les dates ont été modifiées, tout en restant dans la même année. Par exemple, le « 18 novembre 2010 » a pu être changé en « 12 mai 2010 ». Enfin, toutes les adresses sont devenues « 3333 Burnet Ave., Cincinnati, OH 45229 ».

Chaque lettre du corpus d'entraînement et de test a été annotée par au moins trois annotateurs. Il leur était demandé d'identifier les émotions et expressions suivantes : maltraitance, colère, accusation, peur, culpabilité, désespoir, tristesse, pardon, joie, paix, espoir, amour, fierté, reconnaissance, instructions et informations. Une interface Web conçue spécialement a été utilisée pour collecter, gérer et arbitrer l'activité des annotateurs. Cet outil permet des annotations au niveau du mot et de la phrase. Il permet d'annoter un segment avec plusieurs concepts. Cette fonctionnalité a rendu impossible l'utilisation d'un simple coefficient Kappa pour calculer l'accord inter-annotateurs. Celui-ci a donc été calculé à l'aide de l'alpha de Krippendorff (Krippendorff, 1980, 2004), avec une distance de Dice. L'accord moyen mesuré a été de 0,54 (Pestian *et al.*, 2012a).

## 2.2 Recrutement des annotateurs

Par manque de financement, les annotateurs du projet ont été recrutés par *crowdsourcing* bénévole. Le choix a cependant été fait de limiter l'appel à participation à certaines personnes (il s'agit donc de *crowdsourcing* limité), dont le vécu pouvait être vecteur de motivation.

Ainsi, approximativement 1 500 membres de plusieurs communautés en ligne ont reçu une information concernant l'étude, soit directement par courriel, soit indirectement, *via* des pages `Facebook` de soutien. Les deux groupes les plus actifs dans ces communautés ont été les groupes de Karyl Chastain Beal *Families and Friends of Suicides* (Familles et amis de suicidés) et *Parents of Suicides* (Parents de suicidés), ainsi que *Suicide Awareness Voices of Education* (Voix pour l'éducation et la sensibilisation au suicide), un groupe dirigé par Dan Reidenberg, un psychiatre.

Le message envoyé aux participants potentiels contenait des informations concernant l'étude, ses sources de financement et ce qui était attendu des participants. Les volontaires ont été sélectionnés en deux étapes. La première a consisté à vérifier qu'ils remplissaient les critères fixés : avoir au moins 21 ans (la majorité légale aux États-Unis), être de langue maternelle anglaise et être prêt à lire et annoter 50 lettres de suicidés.

Dans un second temps, les participants ont reçu un courriel leur demandant de décrire leur relation à la personne suicidée, le temps passé depuis cette mort et si le ou la suicidé(e) avait été diagnostiqué(e) comme souffrant d'une maladie mentale.

## 2.3 Formation des annotateurs

Les annotateurs ont été formés par le biais d'une interface Web. Ils ont annoté 10 lettres jusqu'à atteindre une qualité minimale de 50 % d'accord observé avec la référence. Ils ont ensuite été invités à annoter 50 autres lettres.

Il est important de noter qu'ils ont été informés de la possibilité d'arrêter quand ils le souhaitaient, durant la formation ou pendant l'annotation. Des possibilités de soutien leur ont été proposées. Un « filet de sécurité » psychologique a été mis en place à deux niveaux : les organisateurs ont proposé un contact auprès de l'équivalent de SOS Suicide et les communautés de provenance des annotateurs leur ont fourni un soutien complémentaire.

## 2.4 Profils des annotateurs

L'annotateur type, dans ce projet, est une femme d'âge moyen, ayant fait des études et dont un membre de la famille proche s'est suicidé.

Le tableau 2 détaille les informations démographiques dont nous disposons sur les annotateurs du projet.

| | | |
|---|---|---|
| Genre | Homme | 10 % |
| | Femme | 90 % |
| Age | Moyenne | 47,3 ans |
| | Écart type | 11,2 ans |
| | Extrêmes | 23–70 ans |
| Niveau d'études | Bac | 26 % |
| | Bac+2 | 13 % |
| | Bac+3 | 23 % |
| | Bac+5 | 34 % |
| | Doctorat | 4 % |
| Connexion au suicide | Survivant d'une perte | 70 % |
| | Prof. de la santé mentale | 18 % |
| | Autre | 12 % |
| Temps passé depuis le suicide | 0-2 ans | 27 % |
| | 3-5 ans | 25 % |
| | 6-10 ans | 14 % |
| | 11-15 ans | 13 % |
| | 16+ ans | 12 % |
| Relation au suicidé | Enfant | 31 % |
| | Frère/soeur | 23 % |
| | Époux ou partenaire | 15 % |
| | Parent | 8 % |
| | Autre parent | 9 % |
| | Ami(e) | 5 % |

TABLE 2 – Démographie des annotateurs

# 3 Résultats

## 3.1 Un corpus annoté unique, disponible pour la recherche

Le corpus est composé de 1 319 lettres, ce qui correspond à un total de 146 739 mots. La longueur moyenne d'une lettre est de 102,4 mots et l'écart type est de 112,2 mots : certaines lettres sont très courtes, d'autres très longues. À notre connaissance, il s'agit de loin du plus gros corpus existant de lettres de suicidés écrites en anglais. À titre de comparaison, (Brevard *et al.*, 1990) présente une étude de 20 lettres de suicidés et 20 lettres liées à une tentative de suicide, soit un total de 40 lettres. Les auteurs de (Joiner *et al.*, 2002) ont utilisé le même nombre de lettres, alors que ceux de (O'Connor & Leenaars, 2004) ont eu accès à 30 lettres en provenance d'Irlande et 30 des États-Unis. En ce qui concerne l'anglais, l'étude la plus large à notre connaissance a concerné 224 lettres de 154 sujets (Ho *et al.*, 1998). En ce qui concerne le français, 482 lettres en français ont été utilisées pour (Volant, 1990).

## 3.2 Des systèmes créés, identifiant les émotions

Une *shared task* a été organisée autour de ce corpus par le centre i2b2 (*Informatics for Integrating Biology and the Bedside*) aux États unis en 2011. Les participants ont mis au point des méthodes de TAL pour annoter les émotions exprimées dans les lettres. Cette *shared task* a regroupé 24 équipes, soit 106 scientifiques, en provenance d'Europe, d'Asie, et d'Amérique du nord. Ceux-ci ont présenté leurs résultats lors d'un atelier à la conférence annuelle de l'American Medical Informatics Association (Pestian *et al.*, 2012b).

Les performances s'étagent de 0,30 à 0,61 de F-mesure. Les actes de cette *shared task* ont été cités près de 60 fois au moment où nous écrivons ces lignes (avril 2015). L'impact de ce corpus en termes d'applications, de produits commerciaux, d'articles de recherche ou de brevets reste cependant inconnu à ce jour.

# 4 Discussion

## 4.1 Des biais probables, difficiles à réduire

Nous avons identifié un certain nombre de biais potentiels dans ce projet. Certains viennent des annotateurs, d'autres des chercheurs.

En ce qui concerne les annotateurs, il est possible que les survivants d'une perte et que les professionnels de la santé mentale influent sur la tâche du fait de différents biais émotionnels et cognitifs.

Ainsi, les survivants pourraient être inconsciemment motivés à trouver moins de preuves de souffrance et plus de joie dans les lettres de leurs parents. Les professionnels de la santé mentale pourraient eux être motivés à trouver davantage ou moins de preuves de l'issue fatale d'une maladie.

Il a été noté dans (Olson, 2005) que les lettres de suicidés peuvent avoir pour but soit de réduire la souffrance des survivants en allégeant leur culpabilité soit, au contraire, à l'alourdir en les accusant. Il est possible que les survivants et les professionnels de la santé mentale annotent les lettres d'adieux à la vie de manière différente, du fait d'une tentative inconsciente des premiers à alléger leur propre souffrance.

Nous n'avons malheureusement pas de réponse à ces questions, mais il est important d'avoir ces questions en tête lorsque l'on conçoit un tel projet d'annotation et lorsque l'on utilise le produit de ce projet.

Ainsi, dans ce projet, des différences ont été observées entre les différents types d'annotateurs. Une analyse préliminaire suggère en effet que les volontaires non professionnels identifient un variété moindre d'émotions que les professionnels de la santé mentale.

> "We conjecture that part of this difference is due to psychological phenomenology. That is, each annotator has a psychological perspective that he/she brings to emotionally-charged data and this phenomenology causes a natural variation. [...] Whether our use of vested volunteers biased the interoperation, we are not sure. " (Pestian *et al.*, 2012b)

Il est également possible que les chercheurs eux-mêmes aient pu inconsciemment biaiser les résultats. Ainsi, un des membres du projet est lui-même un parent proche d'une personne qui s'est suicidée. Il est de ce fait difficile d'identifier clairement les biais introduits dans la conception et l'analyse du projet, mais la question mérite d'être posée.

Quoi qu'il en soit, le travail en équipe participe à limiter ce type d'influence et à protéger le psychisme des uns et des autres, en leur permettant d'échanger sur leur ressenti et d'identifier ainsi plus facilement les biais qui en découlent.

## 4.2 L'empathie en question

Lire et annoter des lettres de suicidés n'est pas une tâche facile. Il est probable que pour la mener à bien les annotateurs doivent être amenés à faire appel à leur empathie, mais également à en réprimer une partie pour se protéger psychologiquement.

En théorie, cela ne devrait pas être différent d'autres situations où l'empathie est nécessaire, comme dans le cas de soins psychiatriques. Ainsi, (Capuzzi & Golden, 2013) souligne la nécessité d'une relation empathique entre le thérapeute et ses patients adolescents et (McLaughlin, 2007) insiste sur le rôle de l'empathie dans les réactions thérapeutiques face au comportement suicidaire. Cependant, dans le cas de cette campagne les annotateurs n'étaient pas formés à cela, ce qui a pu affecter leur capacité à gérer leurs sentiments et les effets secondaires de la campagne. Nous rappelons que les annotateurs pouvaient arrêter n'importe quand s'ils avaient des difficultés (cette option leur était rappelée) et que différentes options leur étaient proposées pour du soutien psychologique.

Les chercheurs qui ont organisé la campagne ont également été touchés, voire bouleversés par leur contact avec ce corpus. De fait, les chercheurs les plus impliqués dans le projet ont l'obligation d'avoir un suivi psychiatrique ou religieux tous les trimestres. Par ailleurs, une rotation régulière du personnel a lieu entre les différents projets.

### 4.3 Des utilisations potentiellement dangereuses des données

Les questions éthiques liées aux applications possibles d'un projet de recherche se posent sans doute dans tout le domaine informatique (Ermann *et al.*, 1997; Friedman, 1997; Martin & Weltz, 1999), et il n'y a aucune raison de penser que le domaine du TAL fasse exception.

Cependant, ce projet est particulier en ce qu'il touche à la mort et au respect de la volonté.

Les conséquences de l'utilisation des technologies issues de ce corpus pourraient être particulièrement sévères, comme l'hospitalisation injustifiée, l'emprisonnement, etc. Il est en effet assez facile d'imaginer l'utilisation qui pourrait être faite de technologies issues de ce corpus par des gouvernements pour lesquels la psychiatrie serait un outil d'oppression des dissidents comme, à une certaine époque, l'Union soviétique, où l'abus de diagnostics psychiatriques était massif (British Medical Association, 1992).

Les buts recherchés eux-mêmes posent question. En effet, est-ce qu'une intervention pour éviter un suicide ne porte pas atteinte au droit de mourir ? C'est une question connue des scientifiques travaillant (ou ayant travaillé) sur le sujet. Ainsi, Joiner remarque que les psychiatres avec qui il a travaillé « respectaient l'autonomie finale des personnes, y compris leur liberté de se donner la mort si c'était ce qu'elles souhaitaient vraiment » [6] (Joiner, 2009). L'un des chercheurs les plus connus du domaine (Edwin Shneidman) dit également «. . .vous me demandez, eh bien, combien de suicides vous voulez, je répondrais que je n'en veux aucun, mais je veux que la liberté de se suicider existe » [7] (Pestian, 2010).

## 5 Conclusion

Si nous considérons le corpus par rapport à la *Charte Éthique et Big Data* [8] (Couillault *et al.*, 2014), nous constatons que les conditions de collecte et de distribution des données pour l'annotation sont en accord avec les suggestions éthiques sous-jacentes. La seule question en suspens est la section traitant des données liées aux contributeurs humains. Les trois sous-sections concernées sont les suivantes :

— *si un consentement a été demandé* Dans notre cas, les annotateurs ont clairement consenti à ce que leurs annotations soient distribuées, mais ce n'est évidemment pas le cas pour les auteurs des lettres.
— *si une trace matérielle existe de ce consentement* Le consentement des annotateurs a été obtenu par le biais d'une interaction électronique, suivie, lorsque le besoin s'en est fait sentir, de conversations téléphoniques ou d'échange de courriels.
— *la nature de l'information fournie afin que le consentement soit éclairé* Les annotateurs ont reçu une information lors de la phase initiale de recrutement par courriel (pour ceux qui ont été recrutés directement), puis au cours de la phase de formation (pour tous les annotateurs).

Nous avons présenté ici un certain nombre de questions éthiques liées à la construction d'un corpus de lettres de suicidés annoté en émotions. D'autres points pourraient encore être analysés, notamment en ce qui concerne l'expérience vécue par les annotateurs.

Nous pensons qu'une des leçons les plus importantes à retenir de ce projet est que les participants en contact avec ce type de matériau doivent avoir accès à différents types de soutien. Ainsi, dans un travail mené actuellement sur les adolescents suicidaires, les transcripteurs se voient proposé un suivi.

Enfin, la *Charte Éthique et Big Data* de ce corpus sera disponible prochainement.

6. "understood people's ultimate autonomy, including their freedom to occasion their own death if they really were committed to doing so".
7. ". . .you say to me, well how many suicides do you want, and I say I don't want any, but I want there to be the freedom to do it"
8. Voir : `http://wiki.ethique-big-data.org`.

## Références


BAUME P., CANTOR C. H. & ROLFE A. (1997). Cybersuicide : the role of interactive suicide notes on the internet. *Crisis : The Journal of Crisis Intervention and Suicide Prevention*, **18**(2), 73.

BREVARD A., LESTER D. & YANG B. (1990). A comparison of suicide notes written by suicide completers and suicide attempters. *Crisis : The Journal of Crisis Intervention and Suicide Prevention*.

BRITISH MEDICAL ASSOCIATION (1992). *Medicine betrayed : The participation of doctors in human rights abuses*. Zed books.

CAPUZZI D. & GOLDEN L. (2013). *Preventing adolescent suicide*. Routledge.

CHAMBERLAIN J., POESIO M. & KRUSCHWITZ U. (2008). Phrase Detectives : a web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics'08)*.

COUILLAULT A., FORT K., ADDA G. & DE MAZANCOURT H. (2014). Evaluating Corpora Documentation with regards to the Ethics and Big Data Charter. In *International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Islande.

COUVET D., DEVICTOR V., JIGUET F. & JULLIARD R. (2011). Scientific contributions of extensive biodiversity monitoring. *C. R. Biologies*, **334**, 370–377.

ERMANN M. D., WILLIAMS M. B. & SHAUF M. S. (1997). *Computers, ethics, and society*. Oxford University Press.

FÉDÉRATION FRANÇAISE DE PSYCHIATRIE (2001). *La crise suicidaire : reconnaître et prendre en charge*. John Libbey Eurotext.

FORT K., ADDA G. & COHEN K. B. (2011). Amazon Mechanical Turk : Gold mine or coal mine ? *Computational Linguistics (editorial)*, **37**(2), 413–420.

FORT K., GUILLAUME B. & CHASTANT H. (2014). Creating Zombilingo, a Game With A Purpose for dependency syntax annotation. In *Gamification for Information Retrieval (GamifIR'14) Workshop*, Amsterdam, Pays-Bas.

FRIEDMAN B. (1997). *Human values and the design of computer technology*. Number 72. Cambridge University Press.

GUPTA N., MARTIN D., HANRAHAN B. V. & O'NEILL J. (2014). Turk-life in india. In *Proceedings of the 18th International Conference on Supporting Group Work*, GROUP '14, p. 1–11, New York, NY, USA : ACM.

HO T., YIP P. S., CHIU C. & HALLIDAY P. (1998). Suicide notes : what do they tell us ? *Acta Psychiatrica Scandinavica*, **98**(6), 467–473.

JOINER T. (2009). *Why people die by suicide*. Harvard University Press.

JOINER T. E., PETTIT J. W., WALKER R. L., VOELZ Z. R., CRUZ J., RUDD M. D. & LESTER D. (2002). Perceived burdensomeness and suicidality : Two studies on the suicide notes of those attempting and those completing suicide. *Journal of Social and Clinical Psychology*, **21**(5), 531–545.

KRIPPENDORFF K. (1980). *Content Analysis : An Introduction to Its Methodology*, chapter 12. Sage : Beverly Hills, CA., USA.

KRIPPENDORFF K. (2004). *Content Analysis : An Introduction to Its Methodology, second edition*, chapter 11. Sage : Thousand Oaks, CA., USA.

LAFOURCADE M. & JOUBERT A. (2008). JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. In *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Lyon, France.

LEENAARS A. A. (1988). *Suicide notes : Predictive clues and patterns*. New York : Human Sciences Press.

MARTIN C. D. & WELTZ E. Y. (1999). From awareness to action : Integrating ethics and social responsibility into the computer science curriculum. *ACM SIGCAS Computers and Society*, **29**(2), 6–14.

MCLAUGHLIN C. (2007). *Suicide-related behaviour : Understanding, caring and therapeutic responses*. John Wiley & Sons.

MERRIAM-WEBSTER (consulté en avril 2015). Définition du terme *Crowdsourcing*.

OLSON L. M. (2005). *The use of suicide notes as an aid for understanding motive in completed suicides*. PhD thesis, Department of Health Promotion and Education, University of Utah.

OSGOOD C. E. & WALKER E. G. (1959). Motivation and language behavior : A content analysis of suicide notes. *The Journal of Abnormal and Social Psychology*, **59**(1), 58.

O'CONNOR R. C. & LEENAARS A. A. (2004). A thematic comparison of suicide notes drawn from northern ireland and the united states. *Current Psychology*, **22**(4), 339–347.

PESTIAN J. (2010). A conversation with edwin shneidman. *Suicide and Life-Threatening Behavior*, **40**(5), 516–523G.

PESTIAN J. P., MATYKIEWICZ P. & LINN-GUST M. (2012a). What's in a note : construction of a suicide note corpus. *Biomedical informatics insights*, **5**, 1.

PESTIAN J. P., MATYKIEWICZ P., LINN-GUST M., SOUTH B., UZUNER O., WIEBE J., COHEN K. B., HURDLE J. & BREW C. (2012b). Sentiment analysis of suicide notes : A shared task. *Biomedical Informatics Insights*, **5**, 3–16.

SAGOT B., FORT K., ADDA G., MARIANI J. & LANG B. (2011). Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé. In *Actes de Traitement Automatique des Langues Naturelles (TALN)*, Montpellier, France. 12 pages.

SCHNEIDMAN E. S. (1981). Suicide notes and tragic lives. *Suicide and Life-Threatening Behavior*.

SHNEIDMAN E. S. (1973). Suicide notes reconsidered. *Psychiatry*, **36**(4), 379–394.

SHNEIDMAN E. S. & FARBEROW N. L. (1957a). *Clues to suicide*, volume 56981. McGraw-Hill Companies.

SHNEIDMAN E. S. & FARBEROW N. L. (1957b). Some comparisons between genuine and simulated suicide notes in terms of mowrer's concepts of discomfort and relief. *The Journal of general psychology*, **56**(2), 251–256.

TUCKMAN J., KLEINER R. J. & LAVELL M. (1959). Emotional content of suicide notes. *American Journal of Psychiatry*, **116**(1), 59–63.

VOLANT É. (1990). *Adieu, la vie... : étude des derniers messages laissés par des suicidés*. Bellarmin.

# Ethique conséquentialiste et traitement automatique des langues : une typologie de facteurs de risques adaptée aux technologies langagières

Anaïs Lefeuvre[1], Jean-Yves Antoine[1,] Willy Allegre[2]
(1) Laboratoire dôInformatique de lôUniversité de Tours, 3 place Jean Jaurès, 41000 Blois
(2) Laboratoire dôElectronique, CMRFF de Kerpape, BP 78, 56275 Ploemeur
anais.lefeuvre@univ-tours.fr, Jean-Yves.Antoine@univ-tours.fr, wallegre@kerpape.mutualite56.fr

**Résumé.** Cet article présente une typologie de facteurs de risques concernant les technologies numériques et plus particulièrement les technologies langagières. Son objectif est dôoffrir une grille dôanalyse pour une évaluation critique des recherches et applications du TALN dans une démarche éthique conséquentialiste.

## Abstract.

**Consequentialist ethics and NLP: a typology of risk factors suitable to language technologies.**
This paper details a typology of risk factors that should concern digital technologies and more specifically NLP. It aims at providing an evaluation grid for an ethical assessment of researches and applications.

**Mots-clés :** Ethique, risque, facteur de risques, vulnérabilité, criticité, risque individuel, risque sociétal, TAL.
**Keywords:** Ethics, risk analysis, risk factor, vulnerability, criticality, individual risk, societal risk, NLP.

## 1 Pour une éthique conséquentialiste des technologies langagières

Lôinformatisation de la société et sa mise en réseau ont le plus souvent suscité des discours enthousiastes sur lôémergence dôune société de la connaissance, le fonctionnement non hiérarchisé dôinternet, vision moderne dôun village global par excellence conduisant à une intelligence et une prise de décision collectives. Ce consensus sociétal est désormais battu en brèche par de multiples travaux en sciences sociales (Jarrige 2014). La critique la plus visible concerne le respect de la vie privée sur les réseaux sociaux, le droit à lôoubli numérique, et plus généralement la question de lôémergence dôune société du contrôle permise par ces techniques. Ces questions surviennent à un moment où le TALN a acquis une maturité suffisante pour permettre par exemple une fouille de données intelligente dans de grands flux dôinformations. Jusquôà récemment, les technologies langagières pouvaient sembler moins concernées par des questionnements éthiques que les biotechnologies ou les nanotechnologies. Lôassociation entre analyse automatique intelligente de la langue, masses de données (*big data*) et informatique ubiquitaire requiert désormais que notre communauté scientifique sôinterroge sur son objet de recherche et sur ses applications.

Les réflexions éthiques en TALN se sont jusquôà présent concentrées surtout sur lôanonymisation des données personnelles dans les corpus. En France, cette question est traitée de longue date par la réglementation (loi *Informatique et Libertés* de 1978), hétéro-régulation normative relevant de décisions étatiques. A lôopposé, lôéthique relève dôune auto-régulation non normative, fondée des choix et jugements de valeurs collectifs. Des comités dôéthique ont progressivement été mis en place. Ils ont avant tout un rôle de recommandation et de conseil, même si lôéthique influence désormais le droit par lôintermédiaire de la jurisprudence. La CERNA (Commission de réflexion sur lôÉthique de la Recherche en sciences et technologies du Numérique dôALLISTENE - ALLIance des Sciences et TEchnologies du NumériquE) joue précisément ce rôle dans le cadre des sciences et technologie du numérique, en se positionnant à lôinterface entre les recherches du domaine et de leurs applications industrielles. Elle a émis un premier jeu de recommandation sur lôéthique de la recherche en robotique (CERNA 2014). Le TALN est toutefois relativement absent de ses activités. Dans cet article, nous aimerions précisément apporter quelques éléments méthodologiques pour la mise en place dôune réflexion éthique en TALN.

Une réflexion éthique peut se baser sur deux principaux courants de pensée contemporains. Dôune part, lôéthique conséquentialiste suit une approche téléologique qui consiste à se focaliser non pas sur des principes mais sur les conséquences de nos actions. Ce point de vue guidait lôutilitarisme de Jeremy Bentham ou John Stuart Mills, pour qui toute action est justifiée par ses effets positifs sur le plus grand nombre. Lôéco-éthique contemporaine cherche au contraire à réduire les effets néfastes de nos actions. Les principes de précaution et de responsabilité, théorisés entre autre par Hans Jonas (1990), relèvent précisément de cette logique.

Dôautre part lôéthique déontologique promeut le respect de principes moraux pour régir nos actions, la réflexion éthique devant porter sur lôétablissement de ces principes. Par exemple, Rawls (1987) et sa théorie de la justice proposent une logique contractualiste entre personnes libres et rationnelles : celle-ci établit des principes de fonctionnement de la société, qui sont jugés éthiques si leur processus amont dôélaboration a été équitable.

La réflexion qui est présentée dans cet article relève dôune étude conséquentialiste des effets négatifs des recherches et applications en TALN. Cette démarche téléologique nous semble en effet la plus adaptée pour engager une prise de conscience dans un champ de connaissances nôayant amorcé que de manière embryonnaire une étude réflexive sur ses pratiques. Une méthodologie dôanalyse en termes de risques, positifs ou négatifs cette fois, existe dans des domaines tels que la finance par exemple. A lôinstar de lôanalyse du risque industriel, nous privilégierons ici une attention aux impacts jugés négatifs dans un objectif dôamélioration des recherches menées en TALN, dont on ne doit bien sûr pas négliger les promesses. En pratique, nous proposons de conduire une analyse de risque telle quôenvisagée dans les autres domaines technologiques, sur les technologies langagières que notre communauté concourre à développer. Cette analyse évaluative nous semble dôautant plus nécessaire que nos sociétés postmodernes ne sont désormais plus régies uniquement par la question du partage des richesses mais également par celle du risque technologique (Beck 2001).

La démarche que nous proposons se place dans une perspective dôamélioration des pratiques de recherche par la mise en place dôune réflexion éthique. Comme toute analyse de risque, elle nécessite lôélaboration de protocoles dôévaluation des effets induits par les technologies que nous produisons. Par la connaissance experte de son domaine dôétude, le chercheur en TALN est idéalement placé pour jouer un rôle de lanceur dôalerte sur des problématiques dont il peut être parfois le seul à percevoir le risque. Cela ne remet aucunement en cause le fait quôà terme, les protocoles dôexpérimentation appropriés devront être conçus en collaboration avec les spécialistes de chaque type dôimpact.

Dans un premier temps, nous allons présenter le cadre classique de lôanalyse de risque et voir dans quelle mesure il pourrait sôappliquer au TALN. Nous nous focaliserons ensuite sur la notion de facteur de risques, que nous mettrons en regard dôune première classification de types de risques qui a été inspirée par nos propres travaux sur lôutilisation des technologies langagières pour lôaide aux personnes handicapées. Cette classification sera ensuite affinée sous la forme dôune typologie à 5 niveaux qui forme une grille dôanalyse pour lôétude du risque en TAL, et potentiellement toute technologie numérique. Cette typologique sera illustrée au regard de diverses applications du TAL.

## 2 Quôest-ce que le risque ?

Historiquement, les premières tentatives de modélisation du risque remontent aux XVII$^{ème}$ et XVIII$^{ème}$ et sont le fait de mathématiciens (Huygens, Bernouilli, Pascal) qui sôinterrogeaient sur lôincertitude liée à la notion de risque. La révolution industrielle va profondément renouveler cette notion : alors que le risque était jusque-là lié dans les consciences aux calamités naturelles par essence inévitables, celles-ci vont laisser la place aux catastrophes et crises environnementales liées au processus industriel. Dans nos sociétés modernes, le risque est donc le fait principal des activités humaines ï dôoù la définition par certains dôune nouvelle ère dénommée *anthropocène* (Crutzen et al. 2007).

Il en résulte lôémergence dôune société réflexive du risque (Beck 2001), où le développement du risque nôest acceptable par la population que dans la mesure où lôon cherche à lôévaluer (rapport bénéfice/risque) et le gérer. Côest dans ce cadre que lôon voit apparaître un ensemble de règlementations et de normes liées à la définition du risque et à sa maîtrise. Le référentiel ISO Guide 73 [10] sur le vocabulaire du risque ne lie pas ce dernier à une menace mais, comme Huygens, à lôincertitude dôun évènement. Le risque y est défini comme *"lôeffet de lôincertitude sur lôatteinte des objectifs"*, une note précisant bien que *"un effet est un écart, positif et/ou négatif, par rapport à une attente"*. Cette définition permet dôintégrer le risque financier et économique. Toutefois, dès quôon en revient au risque industriel, sanitaire ou environnemental, ce sont bien les conséquences néfastes des processus qui sont mises en avant. Ainsi, lôUnion Européenne définit comme suit la notion de risque grave pour la santé publique (Union Européenne 2006) :

- risque : probabilité quôun évènement se produise (on retrouve ici la définition ISO)
- risque potentiel grave pour la santé publique : une situation dans laquelle il existe une forte probabilité pour quôun danger grave provoqué par un médicament (é ) affecte la santé publique
- grave : (é ) signifie un danger qui pourrait entraîner la mort, mettre en danger le patient, nécessiter une hospitalisation, entrainer une invalidité (etc é )

Pour décrire le risque dans cette perspective, nous proposons de suivre un cadre normatif standard, comme par exemple les normes européennes (EN 292-1 et EN 1050) relatives aux risques ayant incidence sur la santé humaine. Dans ce type de démarche, le risque est modélisé comme lôassociation de trois concepts :

- **le facteur de risque**, qui caractérise l'élément ou le processus susceptible de causer un risque, donc d'être la cause d'une situation indésirable. La question que nous nous posons donc est de savoir si les technologies langagières doivent être considérées comme des facteurs de risque. Ceci, en ayant conscience que toute technologie complexe peut constituer a priori une source plurifactorielle de risques variés (Beck 2001).
- la **criticité**, qui combine l'impact du risque (son effet ou sa gravité, pour reprendre le règlement européen détaillé plus haut) avec sa probabilité d'occurrence. La question que l'on se pose ici est l'évaluation de l'impact des technologies que nous développons. Cette évaluation peut être expérimentale (étude statistique sur une population de test) ou subjective et introspective (retour d'expérience d'experts, par exemple).
- la **vulnérabilité**, qui revient d'une part à décrire l'objet du risque, à savoir l'élément qui le subit (ce peut être aussi bien un écosystème, un individu ou l'ensemble de la société), et d'autre part ses conséquences (par exemple, la mort ou l'invalidité dans l'exemple du règlement européen cité précédemment).

A notre connaissance, les technologies langagières ont très rarement fait l'objet d'une analyse de risque ï voir toutefois (Kaplan 2014) pour une exception notable. Afin d'amorcer une telle démarche, nous proposons ici de nous focaliser sur les dimensions de vulnérabilité et de facteur de risque, afin déjà d'identifier quelles applications du TALN peuvent présenter des conséquences potentiellement néfastes. La question de la criticité est beaucoup plus délicate à aborder dans l'immédiat, car elle nécessite la mise en î uvre d'expérimentations ou de suivis utilisateurs qui peuvent être potentiellement très lourds (large cohorte de sujets ou étude longitudinale sur une longue durée). Elle ne saurait donc être abordée que sur les facteurs de risques pour lesquels une vulnérabilité critique est caractérisée.

Cette vulnérabilité devient rapidement cruciale dès lors qu'on touche des personnes déjà fragilisées : c'est par exemple le cas des personnes handicapées, pour lesquelles sont développés des systèmes d'assistance et de suppléance de plus en plus efficaces. Le travail qui est présenté ici a ainsi été initié dans le cadre du RTR (Réseau Thématique Régional) *Risques* de la région Centre s'intéressant à l'impact des aides techniques au handicap. Un des objectifs de ce RTR sera précisément de soutenir des études expérimentales de criticité sur des technologies d'assistance qui pour certaines impliquent des traitements linguistiques. Les travaux du RTR ont conduit à une première esquisse de classification du risque (Antoine et al. 2014) qui sert de base à la typologie arborescente présentée ci-après (§3).

## 3 Une première classification de risques et de leur vulnérabilité

Les deux premiers niveaux de la typologie de risques que nous présentons ici relèvent avant tout de la vulnérabilité (Figure 1). Le premier niveau correspond à l'objet de vulnérabilité en distinguant les atteintes à l'individu de ceux qui relèvent d'une dimension sociétale. Le risque est ensuite sous-catégorisé suivant 5 grandes classes d'impacts identifiés.

On classera en *risque individuel* tout risque dont l'objet de vulnérabilité se limite à l'individu (utilisateur d'un système technique ou autre). Dans le cadre de l'aide technique au handicap, on peut ainsi imaginer que l'objet de vulnérabilité soit par exemple un aidant. Le risque individuel est ensuite découpé en trois classes d'impacts principales :

- Le **risque physique** caractérise une atteinte à l'intégrité physique de l'individu par suite de l'usage d'un système technique. Il affecte le corps d'une personne et correspond à des blessures/dégradations, des traumatismes et/ou des handicaps physiques supplémentaires liés à l'utilisation du système.
- Le **risque cognitif** porte sur une altération dommageable de certaines fonctions cognitives du fait de l'utilisation d'un système technique. Dans le cadre de l'aide au handicap, les thérapeutes sont particulièrement attentifs à ce type de risque, une assistance trop importante pouvant par exemple induire une régression cognitive entraînant une perte d'autonomie en l'absence de dispositif d'assistance. De même, il a été montré que l'usage régulier de moteurs de recherche avait un impact sur nos stratégies de mémorisation à long terme (Sparrow et al. 2011). Nous verrons plus loin que le TAL est directement concerné par ce type de risque.
- Le **risque psychique** se traduit par une perturbation des affects, des réactions ou de la perception de la réalité, qui peut être accentuée par l'usage de nouvelles technologies. On peut par exemple citer le stress induit chez certaines personnes par leur usage. Un exemple relevant du traitement de la parole concerne les synthèses vocales qui utilisent la voix désormais perdue d'un patient (cas de maladies neurodégénératives). On manque encore de recul sur l'introduction de cette pratique, mais nos discussions avec des praticiens montrent que l'adoption de cette voix à forte charge émotionnelle, puisque porteuse forte d'identité, est tout sauf anodine d'un point de vue psychologique.

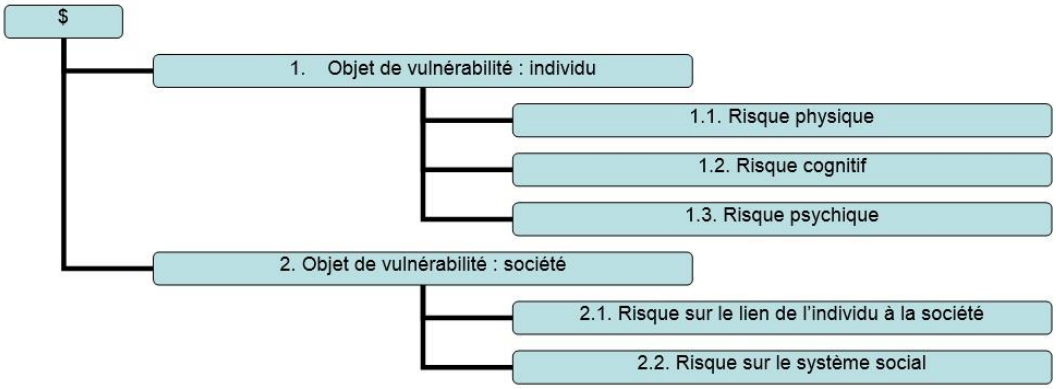


FIGURE 1 : Deux premiers niveau de notre typologie de risques : objet de vulnérabilité et classe d'impact

Le *risque sociétal* met lui en jeu autant les relations sociales des individus que le système social en lui-même. Il s'agit donc d'impacts qui dépassent l'utilisateur seul d'un système techniques, et qui sont souvent sous-évalués. Nous distinguons précisément deux types de risques sociétaux :

- Le risque d'altération des **relations de l'individu à la société** qui concerne aussi bien la modification du lien social que le rapport de l'individu à la société par l'intermédiaire du respect de droits et libertés individuelles. On peut ainsi citer l'impact des technologies numériques sur le droit du travail, le respect de la vie privée, mais aussi les modifications des liens inter-individus avec l'utilisation massive des réseaux sociaux.
- Le risque de modification du **système social**, dans sa dimension politique, économique ou culturelle. Pensons par exemple au vote électronique, ou aux questions d'invisibilité du handicap dans une société numérique.

Cette présentation des premiers niveaux de notre typologique n'a pas fait appel, à dessein, à des exemples relevant des technologies langagières. Nous espérons en effet que cette classification présente un degré de généricité qui lui permet de s'appliquer à d'autres technologies numériques. Nous allons maintenant présenter en détail chaque élément de notre typologie en l'illustrant cette fois avec des applications relevant du TALN, afin de mener une évaluation critique des applications proposées par le TAL dans un contexte où leur diffusion va aller croissante dans les années à venir.

# 4 Une classification typologique des facteurs de risques adaptée au TALN

La classification des facteurs de risques que nous avons introduite dans le paragraphe précédent a été présentée dans le cadre de la journée d'études « Ethique et TAL » organisée par l'ATALA (Antoine et al. 2014). Si elle fournit un cadre général que nous espérons utile pour une analyse du risque du TAL, elle reste toutefois trop générale pour permettre la mise en place d'un diagnostic éthique face à une thématique de recherche précise. C'est pourquoi nous avons continué à affiner cette classification pour proposer une caractérisation du risque suivant une typologie pouvant présenter jusqu'à cinq niveaux hiérarchiques de caractérisation. La typologie que nous proposons pourrait certainement répondre en partie aux interrogations concernant d'autres technologies numériques (big data et analyse décisionnelle, réseaux etcé ). Dans cette section, nous allons toutefois chercher à illustrer notre typologie en nous focalisant uniquement sur le domaine des technologies langagières.

## 4.1 Risque physique

Le risque physique concerne assez peu une technologie comme le traitement des langues qui a rarement prise sur l'environnement physique. Il ne doit toutefois pas être négligé. Notre typologie distingue ici les facteurs de risques concernant l'utilisateur d'un système technique de ceux entrainant un dommage sur l'environnement physique. Dans ce dernier cas, on peut citer l'exemple d'un fauteuil roulant autonome piloté par commande vocale qui heurterait un mur ou un meuble suite à une erreur de commande ou de reconnaissance. Tout facteur de risques correspond à un nî ud terminal de notre typologie. Ici, il s'agit du nî ud 1.1.2 (figure 2).

Un impact physique direct sur l'utilisateur a été caractérisé dans le cas de l'utilisation de la reconnaissance vocale dans les centres logistiques. Afin de permettre un travail mains libres, les préparateurs de commande sont guidés dans leur mission grâce à un dialogue oral homme-machine. Ce mode de gestion entraîne une densification du travail qui peut entraîner une augmentation des lombalgies ou des troubles musculo-squelettiques (INRS 2009). Notre typologie

distingue deux types d'impacts sur l'utilisateur : les atteintes physiologiques internes (modification hormonale par exemple) et les atteintes physiques externes correspondant précisément aux effets néfastes que nous venons d'évoquer.

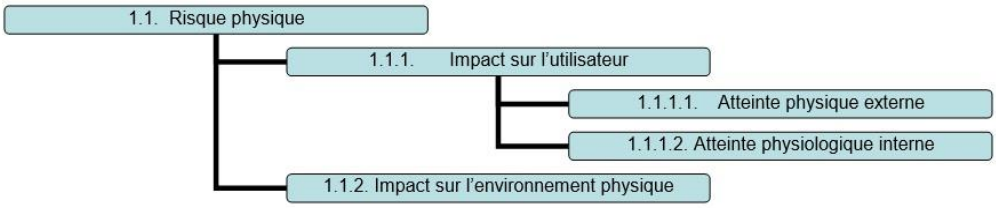


FIGURE 2 : Sous-typologie des risques physiques

### 4.2 Risque cognitif

Bien que notre communauté scientifique se soit peu penchée sur la question par le passé, le risque cognitif concerne de manière significative les technologies langagières. Nous l'avons dit dans la section précédente, ce risque porte sur une altération dommageable de certaines fonctions cognitives ou de l'état cognitif général de l'utilisateur. Comme le montre la figure 3, le premier sous-niveau de classification du risque cognitif concerne précisément cette distinction.

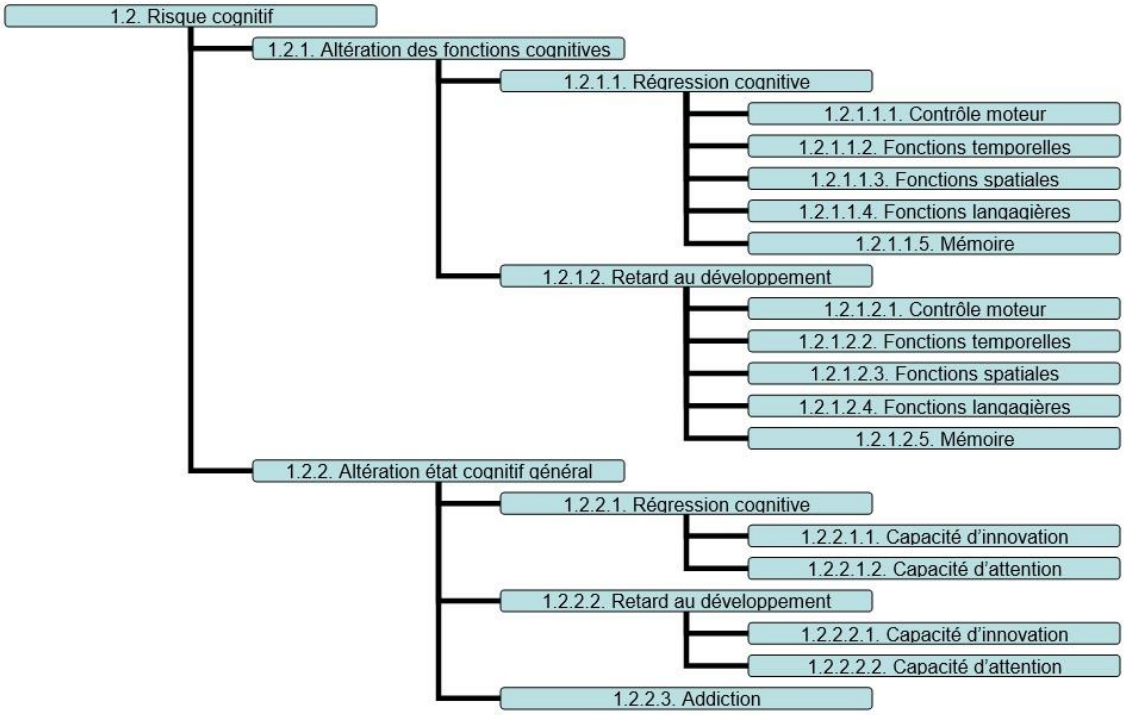


FIGURE 3 : Sous-typologie des risques cognitifs

Le niveau suivant distingue la régression cognitive (déjà vue plus haut), qui entraîne la perte partielle ou totale d'une fonction cognitive du fait de l'usage d'une technologie, du retard au développement qui se traduit par une altération de l'acquisition de cette fonction du fait de l'usage d'une technologie lors des stades de développements cognitifs enfantins. Ces deux types d'altérations sont ensuite différenciés suivant la fonction cognitive concernée :

- Contrôle moteur, mémorisation, fonctions liées aux représentations spatio-temporelles ou au langage, etc. dans le cas de l'altération des fonctions cognitives
- Fonctions plus globales telles que la capacité d'attention ou d'innovation dans le cas des atteintes à l'état cognitif général de l'individu

Cette structuration s'inspire directement de la classification CIF des fonctions mentales globales et spécifiques proposée par l'Organisation Mentale de la Santé (OMS 2001). La CIF peut d'ailleurs servir de cadre à une généralisation ou un raffinement (niveaux hiérarchiques supplémentaires) de notre typologie. S'ajoute enfin la question des addictions cognitives qui ont un impact général sur le fonctionnement cognitif. Toutes ces atteintes ne

concernent pas le TALN. Nous allons toutefois donner quelques exemples de risques liés aux technologies langagières qui montrent lôimportance dôune réflexion sur les impacts cognitifs des technologies que nous élaborons.

**Altération des compétences langagières : prédiction de mots ou auto-complétion pour lôaide à la saisie de texte** ï Ce type dôapplications relève dôune modélisation statistique du langage ou dôune consultation de lexique dans le cas des systèmes les plus simples. Il concerne aussi bien lôaide à la communication pour personnes handicapées que la saisie de texte sur dispositifs mobiles (Antoine 2011). Ces techniques sont généralement évaluées à lôaulne des bénéfices quôelles permettent en termes de vitesse de saisie, voire parfois de développement cognitif. Par exemple, il a été observé que la prédiction lexicale avancée du système dôaide à la communication Sibylle (Wandmacher et al. 2008) induisait un accroissement des productions langagières chez des enfants infirmes moteurs cérébraux, et que lôaugmentation des interactions langagières qui en résultait favorisait le développement cognitif de certains enfants. En outre, leurs enseignants de lôécole intégrée au centre de Kerpape ont remarqué une baisse notable des fautes dôorthographe commises. Une analyse plus fine demande toutefois dôétudier lôimpact de ces technologies dôun point de vue rapport bénéfices/risques. La question que se posent en effet orthophonistes et éducateurs est de savoir si cette aide augmente la maitrise du système de la langue, ou si elle ne masque au contraire pas un abandon de cette capacité au profit du système. Ce risque demande à être évalué en termes de criticité, ce qui demanderait la mise en place de tests de compétences linguistiques longitudinaux. Ce risque se classe sous le nî ud 1.2.1.2.4. de notre typologie, soit suivant le chemin: `Individu > Risque Cognitif > Retard au développement > Fonctions langagières`

Ces aides à la saisie et à la composition de texte se retrouvent dans les correcteurs orthographiques et grammaticaux comme dans toutes les formes dôaide à la traduction désormais disponibles. Qui parmi nous nôa pas ressenti une forme de perte de maîtrise de ses compétences langagières due à lôabandon facile à ces aides techniques ? Pour paraphraser le philosophe Bernard Stiegler (Stiegler 2015), lôhomme augmenté (par les technologies langagières) est un homme diminué (en termes de compétences langagières). Nous sommes ici en présence dôune situation de régression cognitive dont il conviendrait dôévaluer la criticité (impact en termes de compétences langagières).

**Mémoire altérée : texte numérique et abandon de lôécriture cursive** ï Si, comme lôa montré lôexemple précédent, on imagine aisément que lôutilisation dôun système de traitement automatique du langage peut influer sur les fonctions langagières dôun utilisateur, cet impact peut également concerner dôautres fonctions cognitives. Côest le cas en particulier des fonctions de mémorisation, pour lesquelles les exemples dôinfluence des technologies numériques se multiplient. Nous avons déjà cité plus haut le *"Google effect"* relevé par (Sparrow et al. 2011). Cet effet est dôautant plus insidieux quôil nôest pas ressenti par les sujets, qui tendent au contraire à surestimer lôimportance des connaissances mémorisées lorsquôils utilisent un moteur de recherche (Fisher et al. 2015). Certaines études nous conduisent à nous demander si un tel impact négatif ne peut pas être également induit par la saisie numérique de texte au détriment de lôécriture cursive. Suite à lôannonce de lôabandon de lôapprentissage scolaire de lôécriture cursive aux Etats-Unis et en Finlande, de nombreux psychologues du développement ont fait remarquer que la mobilisation des aires motrices lors de lôécriture favorise la mémorisation (Longcamp 2003). La prévalence de lôécriture numérique que favorisent les technologies langagières peut donc induire un risque mémoire dont il serait intéressant dôévaluer la criticité. Ce risque concerne ici encore un éventuel retard au développement classé au nî ud 1.2.1.2.5.

**Capacité dôinnovation et création linguistique : aide à la saisie de texte** ï Les systèmes dôaide à la saisie basés sur la suggestion et lôauto-complétion proposent des items lexicaux, ou lexico-syntaxiques qui sont puisés au sein de ressources linguistiques. Le choix de lôitem suggéré se conforme le plus possible à un idéal linguistique : dans le cas dôun système probabiliste, lôidéal est défini par le nombre, dans le cas dôun système symbolique, côest une norme qui fait loi par exemple. Quelle place accorder dès lors à la créativité linguistique (par le détournement par exemple), et serait-il intéressant de lôévaluer ? Depuis lôinvention de lôécriture jusquôà lôarrivée de la télévision, on assiste à lôépanouissement de telles tendances directives étudiées par la sociolinguistique (voir par exemple à la revue électronique Glottopol pour une approche des politiques linguistiques et lôimpact des média sur les pratiques discursives), il semble donc intéressant de mettre en évidence ce risque que nous avons placé à lôindice 1.2.2.2.1.

### 4.3 Risque psychique

Le risque psychique se traduit par une perturbation de lôétat psychologique de lôindividu de par lôusage dôun système technique. Ces altérations peuvent concerner les affects, les réactions aux situations, la perception de la réalité etc. Leurs effets peuvent être temporaires, comme lors dôun choc émotionnel ou bien plus durables comme dans le cas dôun renforcement dôétat dépressif. Le premier niveau de sous-catégorisation de notre typologie du risque psychique, proposée en figure 4, repose précisément sur une analyse de la vulnérabilité en termes de durée temporelle (durable ou temporaire). Nous distinguons ensuite le risque suivant la variable psychologique qui est affectée chez lôindividu.

Dans le cas dôun impact psychique durable, il nous semble envisageable de nous baser sur les différentes classifications internationales de troubles mentaux (OMS 2006, APA 2003) telles que celles proposées par lôOMS (CIM-10) et lôAssociation Américaine de Psychiatrie (DSM-IV-TR). Bien que parfois controversées, ces classifications sôaccompagnent de critères diagnostiques qui peuvent être utiles pour une étude en criticité dôun facteur de risque. Ces classifications concernent toutefois le plus souvent des formes sévères de troubles psychiques. On peut se demander si les technologies langagières constituent des facteurs de risques dôune telle criticité. Ces classifications doivent donc avant tout être considérées comme des lignes directrices dôanalyse. Notre réflexion sur les technologies langagières et/ou dôaide au handicap nous a toutefois conduits à caractériser des effets potentiels se rapprochant de troubles de lôhumeur (CSM-10 F32 à F38), de troubles anxieux ou de la personnalité (CSM-10 F60).

Nous avons tenu par ailleurs à distinguer des effets temporaires que nous avons observés lors de nos travaux sur lôaide à la communication pour les personnes handicapées : fatigue psychique, situations de stress ou chocs émotionnels. Ce type dôimpact ne doit pas être négligé au titre de son caractère transitoire : par exemple, la fatigue psychique dû à lôeffort cognitif demandé par lôusage de tels systèmes est une cause très fréquente dôabandon de leur usage. Cet impact se retrouve en situation de travail avec le guidage par commande vocale des préparateurs logistiques (INRS 2009).

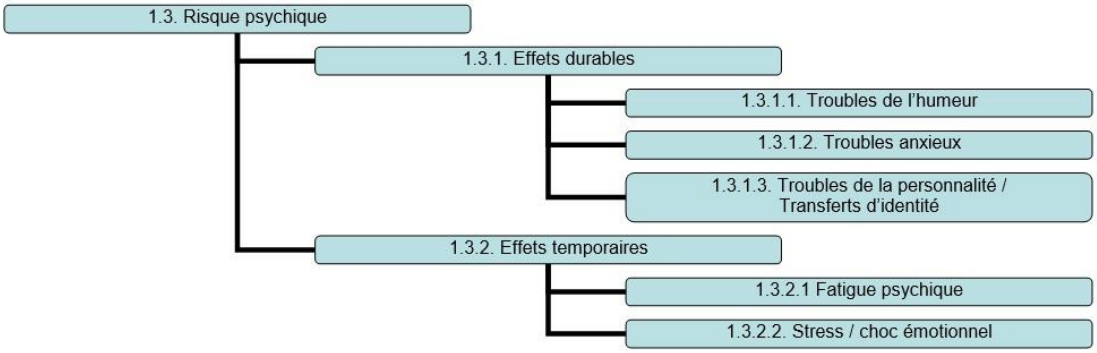


FIGURE 4 : Sous-typologie des risques psychiques

Quelques exemples vont nous permettre dôillustrer lôapplication de cette sous-typologie sur des applications relevant du traitement des langues naturelles ou du traitement de la parole.

**Synthèse vocale, Agents Communicationnels Artificiels (ACA) : choc émotionnel et troubles de lôidentité** ï La synthèse vocale à partir du texte a atteint désormais un degré de naturalité suffisante pour permettre des applications grand public de plus en plus confondantes. On la retrouve ainsi pour les annonces dans certains réseaux de transports en commun. Côest son utilisation à des fins personnalisées qui interroge toutefois en priorité. On pense aux synthèses vocales qui utilisent la voix désormais perdue dôun patient dans le cas de maladies neurodégénératives. Nos discussions avec des praticiens montrent que lôadoption de cette voix à forte charge émotionnelle, puisque porteuse dôidentité, est tout sauf anodine dôun point de vue psychique. Deux situations semblent être à contrôler avec attention :

– lôannonce au patient quôil va perdre sa voix et quôil faut sôy préparer en procédant à son enregistrement : cette séance de recueil peut induire un choc émotionnel critique que la technologie peut renforcer (risque 1.3.2.2).
– lôusage ultérieur de la voix recueillie avec un logiciel dôaide à la communication, qui peut entraîner des effets durables en termes dôidentité (risque 1.3.1.3). Les premiers retours dôexpérience suggèrent que ce facteur de risques concerne, en termes de vulnérabilité, la personne handicapée comme son entourage (famille, aidants).

Ces troubles psychiques liés à la synthèse de parole ne peuvent être ignorés : ils nous semblent équivalents à ceux, bien documentés, liés aux greffes dôorganes (Triffaux et al. 2002). Et ce dôautant quôils concernent un trait de personnalité aussi important que le visage, dont la greffe fait lôobjet de questionnements éthiques (Colas-Benayoun et al. 2006).

Ce risque psychique peut être généralisé à toutes tentatives dôimitation du vivant, parmi lesquels les travaux relevant de lôinteraction affective. Dans le domaine du TALN, ils concernent par exemple la détection des émotions ainsi que les recherches en dialogue homme-machine sur les agents conversationnels animés (ACA). Si les ACA destinés au grand public restent encore assez frustres (voir par exemple Laura, du site EDF particuliers : www.bleuciel.edf.com), leur inspiration reste anthropocentrée. On peut sôinterroger sur lôimpact psychique que pourrait avoir à lôavenir lôéchange avec un agent artificiel plus convaincant (phénomènes de transferts, attentes trop fortes dans lôinteraction, etc.). Sur un sujet proche, la CERNA vient précisément de produire des recommandations portant sur la pertinence de lôimitation du vivant dans le domaine de la robotique (CERNA 2014). Ses propositions sont proches de nos préoccupations, comme le montre par exemple la recommandation IVI-1 sur lôutilité au regard des finalités :

*(é ) Dans les cas où lôapparence ou la voix humaines sont imitées, le chercheur sôinterrogera sur les effets que pourrait avoir cette imitation, y compris hors des usages pour lesquels le robot est conçu.*

### 4.4 Risque sur le lien social et sur le rapport de lôindividu à la société

Du point de vue des applications grand public, côest ainsi plus les situations de vulnérabilité sociétale quôindividuelles qui ont retenu lôattention des chercheurs en TALN. Cette réflexion sôinscrit avant tout dans les interrogations contemporaines sur la société de lôinformation numérique créée par Internet : la question de lôanonymisation des corpus, qui relève de fait de la loi Informatique et Liberté du 6 janvier 1978, revêt par exemple un caractère encore plus sensible avec les recherches portant sur les réseaux sociaux.

Toutefois, lôimpact sociétal du TALN dans notre société numérique couvre bien dôautres dimensions qui sont le plus souvent ignorées. La première classe de risques que nous avons tenu à identifier concerne lôaltération des rapports que les individus entretiennent dans la société : lien social inter-individu médié par les outils numériques, relation avec une technologie de plus en plus anthropocentrée, et enfin respect des droits et libertés individuelles que la société définit par lôintermédiaire de la réglementation (droit du travail, propriété intellectuelle, respect de la vie privée, liberté dôexpression et dôinformation, etc.). La sous-typologie de risques qui en résulte est décrite figure 5. Nous allons lôillustrer à lôaide de quelques exemples dôapplications langagières nécessitant analyse.

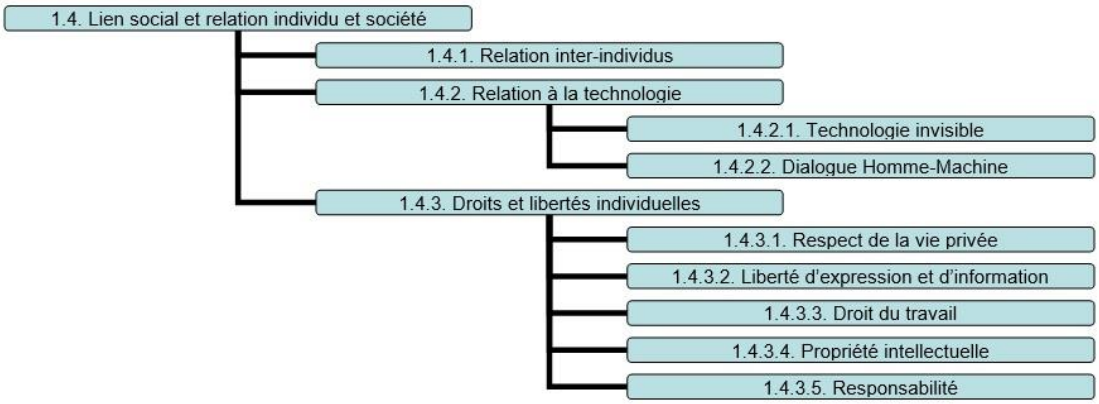


FIGURE 5 : Sous-typologie des risques sur le lien social et le rapport de lôindividu à la société

**Synthèse de parole, agents conversationnel et interaction affective : altération du lien social** ï Nous avons identifié plus haut les risques psychiques que présentent la tentation dôutiliser les technologies langagières et/ou robotiques à des fins dôimitation du vivant. Lorsquôun agent virtuel a pour finalité de se substituer à un être humain dans lôéchange, il existe un risque potentiel de voir lôutilisateur sôenfermer dans des relations purement virtuelles ou fortement médiées par la technologie dans lesquelles la conscience de la distinction entre humanité et intelligence artificielle peut être brouillée. Ici encore, ce risque a été étudié par la CERNA, dont nous rappelons les recommandations (CERNA 2014) :

– *Si une ressemblance quasi parfaite est visée, le chercheur doit avoir conscience que la démarche biomimétique peut brouiller la frontière entre un être vivant et un artefact* (recommandation IVI-2).
– *Pour les projets (é ) qui ont trait au développement de la robotique affective, le chercheur sôinterrogera sur les répercussions éventuelles de son travail sur les capacités de socialisation de lôutilisateur* (IVI-3)

On observe que ces impacts potentiellement négatifs concernent aussi bien lôindividu (risque 1.3.1.3 évoqué plus haut) que son rapport à autrui : ce second risque est classifié 1.4.1 dans notre typologie. Au fur et à mesure des progrès futurs de lôinteraction homme-machine, il est probable quôun risque spécifique devra être identifié, tant une relation dôordre nouveau devra être identifiée. Nous lôavons classifié au nî ud 1.4.2.2 de notre typologie.

**Synthèse vocale, génération automatique de texte : invisibilité de la technologie** ï Les exemples précédents de brouillage entre être humain et artefact concernent lôinteraction directe entre individus. Dôune manière plus générale, il est important de sôinterroger sur notre rapport à la technologie. Quelle place devons-nous lui accorder dans nos sociétés techniques ? Dans quelle mesure redéfinit-elle la notion dôHumanité ? Pour aborder ces questions, lôutilisateur doit avoir une conscience claire des interventions de la technologie dans son quotidien. Cette question de lôinvisibilité de la technologie concerne la synthèse vocale, déjà évoquée. Il faut également évoquer lôintervention de la génération

automatique de texte dans des domaines aussi variés que la gestion automatique de la relation client et la rédaction automatique de compte-rendus boursiers ou de résultats électoraux. Ces interventions sont si sensibles que le quotidien *Le Monde* a tenu à les expliciter à lôoccasion de la dernière consultation départementale (Le Monde 2015). Nous classons ces problèmes dôinvisibilité sous le nî ud 1.4.2.1 de notre typologie.

**Détection dôauteur sur les réseaux sociaux : respect de la vie privée** ï Le respect de la vie privée est certainement un des droits individuels sur lesquels la population est la plus sensibilisée avec lôémergence des réseaux sociaux. Cela nôempêche toutefois pas de nombreux chercheurs en TALN et recherche dôinformation de mener des recherches sur la recherche dôauteurs non identifiés explicitement (Stamatatos 2009). Historiquement, ces recherches en AA (*Authorship Authentification*), issues de la stylistique quantitative, ont eu une justification scientifique (attribution dôauteur à des manuscrits littéraires anciens) puis juridique (détection de plagiat). Leur application à la fouille sur réseaux sociaux pose au contraire des problèmes éthiques qui recommanderaient de sôinterroger sur lôimpact de ces recherches : doit-on pouvoir lever lôanonymat sur un post publié sous un pseudonyme ? Est-il légitime dôutiliser comme preuve juridique ces techniques à la précision perfectible ? Ces atteintes à la vie privée sont classées au nî ud 1.4.3.1 de la typologie.

**Myriadisation et TALN : droit du travail, propriété intellectuelle** ï Le droit du travail (nî ud 1.4.3.3. de notre typologie) est également une dimension qui ne doit pas être oubliée dans le cadre dôune réflexion éthique. Ici, la technologie nôentre dans la réflexion quôà la marge : depuis la révolution industrielle, le statut de la machine dans lôactivité a été réglementé par la loi. Du fait de lôémergence dôun TAL centré sur les données faisant un large usage de ressources langagières, côest sur nos pratiques quotidiennes que nous devons nous interroger. Le recours à un travail parcellisé et myriadisé (*microsourcing* et *crowdsourcing* en anglais) pour lôobtention de telles ressources doit ici être questionné. Dans leur analyse critique dôAmazon Mechanical Turk, (Sagot et al. 2011) montre ainsi quôune part non négligeable des Turkers y réalise une véritable activité dissimulée représentant une part significative de leurs revenus. Ce travail dissimulé concerne directement le TAL, puisque les tâches de transcription et de traduction y sont fréquentes. Cette question concerne également les GWAP (*Game With A Purpose*) dédiés à la production de ressources linguistiques tels que JeuxDeMots (Lafourcade et Lebrun, 2014) et Zombilingo (Fort et. al. 2014), dont les concepteurs insistent sur lôimportance que lôactivité des joueurs reste strictement ludique et par conséquent non rémunérée.

Ces activités et plus généralement la constitution de toute ressource linguistique posent en outre des questions de propriété intellectuelle identifiées comme le nî ud 1.4.3.4 de notre typologie. Si la législation sécurise ces questions, il est essentiel dôadopter de bonnes pratiques sur la documentation des licences, lôidentification des propriétaires et des intervenants dans la constitution pour pouvoir sécuriser ces dernières (Couillault et Fort 2013, Baude et al. 2006).

**Agents virtuels communicants : responsabilité individuelle** ï Le statut des machines considérées doit être réévalué suite à lôémergence du TAL, dôune part à cause de son impact sur les évolutions des conditions de travail tel que décrit dans la section précédente, mais aussi par lôeffet des productions langagières de ces machines. Dans une société où le lien social est lui aussi mécanisé par des plateformes de mise en relation, lôintroduction dôagents conversationnels autonomes est une pratique déjà très répandue[1] et une véritable économie se développe sans que les questions de responsabilités quôelles induisent nôaient été discutées. Les actes performatifs nuisibles tels que la diffamation, lôinsulte, etc. sont déjà encadrés dans la sphère publique, mais quel statut juridique doit être accordé à un agent conversationnel personnel à la source de telles données ? Si la jurisprudence ou la loi attribue la responsabilité à lôagent, alors le créateur est dédouané de responsabilité, ce qui fait des agents une arme performative redoutable, tandis que si le créateur est tenu responsable, alors il est possible quôil soit injustement attaqué pour simple mauvais réglages. Se pose dès lors aussi la question de la répartition équitable de la responsabilité dans le cas où lôagent a été créé par plusieurs personnes. Ces situations nourrissent un débat juridique actuel sur la nécessité ou non de lôattribution dôune personnalité juridique aux robots et agents conversationnels (Robolaw 2014, Bensoussan 2015). Elles relèvent du risque 1.3.4.5 de notre typologie.

### 4.5 Risque sur le système social

Enfin, le dernier type de risques concerne la modification de lôétat de la société dans sa globalité et dans au moins une de ses composantes. Nous pensons ainsi à lôinfluence des technologies numériques sur la prise de décision politique. Ainsi, les aides techniques aux personnes handicapées peuvent limiter leur visibilité dans lôespace politique, avec pour conséquence éventuelle une limitation des politiques dôaides à leur égard et à celle des aidants dans nos sociétés

[1] Le nombre de « bots » utilisés pour la publicité sur les plateformes de rencontres a décuplé ces derniers temps, et même si le test de Turing n'est pas encore passé, le « bot » Ava a réussi le test de Tinder (http://www.adweek.com/adfreak/tinder-users-sxsw-are-falling-woman-shes-not-what-she-appears-163486).

budgétairement contraintes. La question de l'influence de l'automatisation sur le marché de l'emploi et l'économie est documentée depuis les débuts de la révolution industrielle. Nous verrons dans les exemples ci-dessous que cette question ne peut être éludée par les technologies langagières. Enfin, l'impact socio-culturel du TALN associé aux grandes masses de données disponibles sur les réseaux, ne peut plus être ignoré. En particulier, il convient de s'interroger sur l'influence des technologies langagières sur le système même de la langue. La figure 6 ci-dessous résume ces différents types de risques que nous allons une nouvelle fois illustrer à l'aide de quelques exemples.


FIGURE 6 : Sous-typologie des risques sur le système social dans sa globalité

**Indexation automatique, fouille de texte et recherche d'information : risque politique et surabondance de l'information** ï Au sein de notre société numérique connectée, les technologies langagières ne sont généralement pas des facteurs de risques primaires, mais des facteurs aggravant de criticité auxquels il convient de prendre garde. Face à une évolution rapide d'une rareté vers une surabondance nocive de l'information (Ganascia 2013 ; Mariani 2013 :14), le TALN a un rôle important à jouer : l'indexation automatique de contenus, la fouille de texte et la recherche d'information peuvent aider l'utilisateur à donner du sens à son immersion dans de larges flux d'information et de connaissances. Mais elles peuvent également participer à l'illusion trompeuse d'une maîtrise de ces grandes masses de données, comme le suggèrent les expérimentations (cf § 4.2) sur le sentiment erroné de connaissance que développent les utilisateurs de moteurs de recherche (Fisher et al. 2015). Ainsi, la mise en ï uvre de robots logiciels pour la rédaction de textes journalistiques, déjà citée, où bien de moteurs de filtrage automatiques nous semblent jouer contre un contrôle conscient de nos stratégies personnelles de contrôle et de sélection de l'information. Ce risque politique est catégorisé au nï ud 1.5.1. de notre typologie.

**Technologies langagières et économie** ï Les technologies langagières sont de plus en plus concernées par l'économie de la société numérique de l'information. Leur impact économique ne peut donc être ignoré. A titre d'exemple (Sagot et al. 2011) fait une revue des études qui ont été menées sur le coût des ressources créées avec Amazon Mechanical Turk. Dans un autre contexte, la recherche d'information a créé "*la bourse des mots*", algorithme permettant à un annonceur de proposer une enchère sur un mot tapé dans un moteur de recherche afin de lui affecter la publicité la plus appropriée (a le plus de chance d'être cliquée et a le plus de chance d'être tapée en premier lieu). Cette économie linguistique très étudiée par Kaplan (2014) conforte une disparité entre langues richement et faiblement dotées. Ce risque (au nï ud 1.5.2 de notre classification) est abordé par d'autres auteurs tels qu'Enguehard et al. (2014).

**Technologies langagières et système de la langue** ï L'usage des technologies langagières influence l'évolution de la langue, tout comme le fit jadis le passage de l'oral à l'écrit : d'un certain point de vue, à chaque bond technologique en matière de supports[2], le système linguistique dans son ensemble est impacté. Par exemple la sémantique et les mécanismes référentiels à l'ï uvre dans l'usage de "*technomorphèmes*" tels que les hyperliens, les hashtags dans les discours numériques commencent à susciter l'intérêt des analystes du discours (Paveau 2014). Fréderic Kaplan (2014) montre par ailleurs que le choix de l'anglais comme langue pivot entre deux autres idiomes par Google Translate se traduit tout d'abord par un biais culturel qu'il faudrait étudier. Mais qu'en outre, les textes produits ainsi automatiquement (et leurs erreurs), peuvent être prises à tort comme ressources primaires et participer à une forme de "*créolisation numérique*". Au-delà d'un simple enrichissement lexical nécessaire à désigner une nouvelle réalité, on assiste à un genre d'hybridation de la langue entre productions naturelles et productions artificielles pour lesquels l'appareillage méthodologique de l'analyste se doit d'évoluer. De même, on pourrait examiner l'impact de la simplification de texte, domaine du TAL qui propose de remédier provisoirement aux difficultés lexicales d'apprenants ou de souffrants de pathologies dans le but de les accompagner vers une compétence langagière de meilleure qualité. Son principe est de remplacer au sein d'un texte un item lexical par exemple dont la complexité est jugée trop poussée, par un concurrent sur son axe paradigmatique dont la complexité est inférieure. Cette méthode permettrait de rendre accessibles à l'utilisateur diverses productions discursives dont des ï uvres littéraires par exemple. On peut chercher à évaluer l'impact de cette pratique sur l'intégrité et la portée des discours, et à terme sur l'évolution du système linguistique dans son ensemble. Ces modifications sont regroupées sous le nï ud 1.5.3.

[2] On extrapole un peu ici l'acception de technologies du TAL pour considérer tout l'appareillage numérique utile à la production linguistique : SMS, tweet, forum, etc.

# 5 Conclusion

Dans cet article, nous avons cherché à mener une réflexion éthique sur les conséquences de la diffusion de plus en plus importante des technologies langagières. Cette réflexion s'articule avec une volonté d'évaluation des systèmes développés par notre domaine de recherche. Les technologies langagières ouvrent des possibilités d'augmentation de l'humain de plus en plus présentes dans le quotidien, certaines étant directement issues d'une recherche d'aide aux personnes handicapées (auto-complétion par exemple). Cette situation nous plonge directement dans les questionnements autour du posthumanisme et du transhumanisme, ayant respectivement pour principe la réparation puis l'augmentation de l'humain (Kleinpeter 2013). En ce qui concerne le langage, nous tenons à rappeler que les technologies permettant une augmentation de la compétence langagière devrait être mises en perspective avec celles qui les ont précédées, à savoir l'écriture, l'imprimerie, ou encore la télévisioné C'est donc dans cette perspective que nous cherchons à savoir en quoi l'augmentation proposée par ces outils reste une forme d'amélioration et à limiter les effets pervers de l'utilisation massive de ces derniers par une maîtrise étroite des risques qui lui sont liés.

Ce que nous avons tenté de montrer est que, comme toute autre technologie, les technologies langagières, ne sont pas des objets neutres mais ont un impact individuel et sociétal sur lesquels le chercheur doit s'interroger. Une pratique éthique de recherche en TALN ne peut donc se résumer aux questions importantes d'anonymisation qui ont le plus souvent concentré l'attention de la communauté, mais concerner plus généralement les risques psychologiques, cognitifs, sociétaux induits par ces technologies.

Une conclusion préliminaire à ce travail est la nécessité de classer le facteur de risques d'une part et de classer les impacts d'autre part. Nous avons ainsi proposé une typologie de risques qui permet de guider une évaluation éthique des technologies langagières. Nous avons vu qu'un facteur de risquesp eut impliquer plusieurs impacts de vulnérabilité de notre typologie (des capacités cognitives d'un locuteur au système linguistique dans son ensemble). Cette analyse des facteurs de risques et de leur vulnérabilité associée doit maintenant être complétée, quand cela le nécessite, d'une étude expérimentale de leur criticité. C'est ce que nous nous envisageons de faire désormais dans le cas des technologies langagières d'aide au handicap, ceci dans le cadre du RTR Risques de la région Centre.

## Remerciements

Nous tenons à remercier Christian Toinard (LIFO, U. Orléans) pour sa participation à la réflexion sur les risques induits dans le domaine de l'aide au handicap et Adrien Granger pour avoir partagé son expertise en matière de "*chatbots*" et ses réflexions sur la responsabilité engagée par ceux-ci (§ 4.4, sous-section sur les agents virtuels).

## Références

ANTOINE J-Y. (2011) Prédiction de mots et saisie de requêtes sur interfaces limitées : dispositifs mobiles et aide au handicap, In. Bellot P. (Ed). *Recherche d'information contextuelle, assistée et personnalisée*. Hermès, Paris. 273-298.

ANTOINE J.Y., LEFEUVRE A., ALLEGRE W. (2014). Pour une réflexion éthique sur les conséquences de l'usage des NTIC: le cas des aides techniques (à composante langagière ou non) aux personnes handicapées. *Journée ATALA "Ethique et TAL"*, novembre 2014.

ANTOINE J.Y., LABAT M-E., LEFEUVRE A., TOINARD C. (2014b) Vers une méthode de maîtrise des risques dans l'informatisation de l'aide au handicap. Actes *Enviroriskô2014*. Bourges.

APA (2003) *DSM-IV-TR, Manuel diagnostique et statistique des troubles mentaux*, Elsevier Masson, Paris 2003.

BAUDE O. ET AL (2006) *Corpus oraux, guide des bonnes pratiques*. Presses Universitaires d'Orléans, CNRS Editions.

BECK U.(2001). *La société du risque*, Flammarion, Champs/essais.

BENSOUSSAN A. (2015) Faut-il des lois pour nous protéger des robots ? *L'Expansion*. 4 avril 2015.

CERNA.(2014). Ethique de la recherche en robotique. Rapport n° 1 de la CERNA. CERNA ï ALLISTENE . Section IVI : l'imitation du vivant et l'interaction affective et sociale avec les humains.

COLAS-BENAYOUN M.D., FIDELE G., FAVRE J.D. (2006), De la défiguration à la transfiguration : la greffe d'un visage est-elle la solution ? *Annales Médico-Psychologiques*, 16(8), 687-691.

COUILLAULT A., FORT K. (2013) Charte éthique et big data : parce que mon corpus le vaut bien ! Acte colloque *Linguistique, Langue et Parole : statuts, usages et mésusages*. Strasbourg.

CRUTZEN P.J., STEPHEN W., MC NEILL J. (2007), The Anthropocene: are Humans now Overwhelming the Great Forces of Nature? *Ambio*, 36, 614-621.

ENGUEHARD C, MANGEOR M. (2014). Favorisons la diversité linguistique en TAL. *Journée ATALA "Ethique et TAL"*.

FISHER, M., GODDU, M. K., & KEIL, F. C. (2015). Searching for Explanations: How the Internet Inflates Estimates of Internal Knowledge. *Journal of Experimental Psychology*: General. Advance online publication (30 mars 2015) consultee sur : http://dx.doi.org/10.1037/xge0000070.

FORT K., GUILLAUME B., CHASTANT H. (2014) Creating Zombilingo, a Game With A Purpose for dependency syntax annotation. Prof. *Gamification for Information Retrieval Workshop* (*GamifIR'14*), Amsterdam, Pays-Bas.

GANASCIA J.G. (2013) L'initiative Onlife de la commission européenne. Audition auprès de la CERNA-ALLISTENE, 18 mars 2013. Consulté le 30/03/2015 sur : http://cerna-ethics-allistene.org/digitalAssets/31/31320_Onlife.pdf

INRS (2009) *Fiche pratique de sécurité ED 135. Préparation de commande guidée par reconnaissance vocale*.

ISO (2009) ISO Guide 73:2009(fr): management du risque : vocabulaire. Consulté le 20/11/2014 sur https://www.iso.org/obp/ui/fr/#iso:std:44651:fr

JARRIGE F. (2014) *Technocritiques : du refus des machines à la contestation des technosciences*. La Découverte.

JONAS H. (1990) *Le principe responsabilité*. Le Cerf, Paris.

KAPLAN F. (2014) Linguistic Capitalism and Algorithmic Mediation. *Representations* 127 (1): 57–63.

KLEINPETER E. (Dir). (2013) *L'humain augmenté*. CNRS Editions, coll. *"Les essentiels d'Hermès"*.

LAFOURCADE M., LEBRUN A. (2014) Ethique et construction collaborative de données lexicales par des GWAPs (quelques leçons tirées de l'expérience JeuxDeMots). Actes journée d'étude *"Éthique et TAL"* de l'ATALA. Paris.

LE MONDE (2015) Des robots au « Monde » pendant les élections départementales ? Consulté le 24/03/15 : http://makingof.blog.lemonde.fr/2015/03/23/des-robots-au-monde-pendant-les-elections-departementales-oui-et-non/

LONGCAMP A. (2003) Etude comportementale et neuro-fonctionnelle des interactions perceptivo-motrices dans la perception visuelle de lettres. Notre manière d'écrire influence-t-elle notre manière de lire? Thèse U. Aix-Marseille II.

MARIANI J. (2013) Pour une éthique de la Recherche en Sciences et Technologies de l'Information et de la Communication. Consulté le 30/03/2015 sur : http://www.lina.univ-nantes.fr/IMG/pdf/COMETS_Mariani.pdf

OMS (2006) *Classification statistique internationale des maladies et des problèmes de santé connexes/ International Statistical Classification of Diseases and Related Health Problems*. Organisation Mondiale de la Santé. Chapitre 5 : troubles mentaux et du comportement.

OMS (2001) *Classification internationale du fonctionnement, du handicap et de la santé*. Organisation Mondiale de la Santé. Chapitre 1 : fonctions mentales. Consultée le 2/04/2015 sur : http://apps.who.int/classifications/icfbrowser/

PAVEAU M.-A. (2014) « L'alternative quantitatif/qualitatif à l'épreuve des univers discursifs numériques », *Corela* [En ligne], HS-15 | 2014, mis en ligne le 15 octobre 2014, consulté le 17/042015 sur : http://corela.revues.org/3598

RAWLS J. (1987) *Théorie de la justice*. Le Seuil, Paris.

ROBOLAW Project (2014) *D6.2 Guidelines for Regulating Robotics*, consulté le 20/05/2015 sur http://www.robolaw.eu/

SPARROW B., LIU J., WEGNER D.M. (2011) Google effects on memory: cognitive consequences of having information at our fingertips. *Science*, 333, 776-778

STATAMATOS E. (2009) A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538-556.

STIEGLER B. (2015) S'augmenter ou se diminuer ? *Entretien BiTS S02E20 : les promesses du transhumanisme sont-elles réalisables ?* ARTE TV, magazine BiTS. Diffusé le 12 mars 2015. http://creative.arte.tv/fr/bits-trans-human

TRIFFAUX J.M., MAURETTE J.L., DOZOT J.P., BERTRAND J. (2002) *Troubles psychiques liés aux greffes d'organes*. Editions Scientifiques et Médicales. Elsevier. Postprint consulté le 30/03/2015 sur : http://hdl.handle.net/2268/80452

UNION EUROPEENNE (2006) Ligne directrice concernant la définition d'un risque potentiel grave pour la santé humaine ou animale ou pour l'environnement dans le cadre de l'article 33, paragraphes 1 et 2, de la directive 2001/82/CE. *Journal officiel de l'Union Européenne*, C133/6. 8.6.2006.

WANDMACHER T., ANTOINE J-Y., DEPARTE J-P., POIRIER F. (2008) Sibylle, an assistive communication system adapting to the context and its user. *ACM Transactions on Accessible Computing*. 1(1). pp. 1-30.