Enjeux méthodologiques, linguistiques et informatiques pour le traitement du français écrit des sourds

Tristan Vanrullen[1] *Leïla Boutora*[2] *Jean Dagron*[3]
(1) TVSI, 13009 Marseille
(2) LPL, UMR 7309 CNRS/Univ. d'Aix-Marseille, 13100 Aix-en-Provence
(3) Assistance publique - Hôpitaux de Marseille, 13005 Marseille
tristan.vanrullen@gmail.com, leila.boutora@lpl-aix.fr,
jean.dagron@ap-hm.fr

RESUME

L'ouverture du Centre National de Réception des Appels d'Urgence (CNRAU) accessible aux sourds et malentendants fait émerger des questions linguistiques qui portent sur le français écrit des sourds, et des questions informatiques dans le domaine du traitement automatique du langage naturel. Le français écrit des sourds, pratiqué par une population hétérogène, comporte des spécificités morpho-syntaxiques et morpho-lexicales qui peuvent rendre problématique la communication écrite entre les personnes sourdes appelantes et les agents du CNRAU. Un premier corpus de français écrit sourd élicité avec mise en situation d'urgence (FAX-ESSU) a été recueilli dans la perspective de proposer des solutions TAL et linguistiques aux agents du CNRAU dans le cadre de ces échanges écrits. Nous présentons une première étude lexicale, morphosyntaxique et syntaxique de ce corpus reposant en partie sur une chaîne de traitement automatique, afin de valider les phénomènes linguistiques décrits dans la littérature et d'enrichir la connaissance du français écrit des sourds.

ABSTRACT

Methodological, linguistic and computational challenges for processing written French of deaf people

With the setup of a national emergency call-center for deaf people in France (CNRAU), some questions arise in linguistics and natural language processing about the written expression of deaf people. It is practiced by an heterogeneous population and shows morpho-syntactic, lexical and syntactic specificities which increase the difficulty, over the emergency situation, to successfully communicate between the deaf callers and the call-center operators. A first corpus (FAX-ESSU) of written French of deaf people was built with emergency conditions in order to provide linguistic and NLP solutions to the call center operators. On this corpus, we present a first study realized with the help of a natural language processing toolbox, in order to validation linguistic phenomenons described in the scientific literature and to enrich the knowledge of written French of deaf people.

MOTS-CLES : Français écrit des sourds, TAL, Français Langue Etrangère, linguistique de corpus, lexique, syntaxe, méthodologie.
KEYWORDS : Written French of deaf people, NLP, French as a foreign language, corpus linguistics, lexicon, syntax, methodology.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 559–566,
Grenoble, 4 au 8 juin 2012.

1 Contexte et motivations de l'étude

1.1 Contexte institutionnel, sociolinguistique et technologique

L'ouverture en septembre 2011 d'un Centre National de Réception des Appels d'Urgence (CNRAU, numéro d'urgence 114) pour les personnes sourdes et malentendantes résulte de l'application du décret publié le 16 avril 2008 (prévu par l'article 78 de la loi du 11 février 2005). Le centre d'appels concerne trois types d'urgence : police, pompiers, SAMU. Il vise à rendre accessibles dans une modalité autre qu'audio-vocale les services d'appels d'urgence à une population spécifique, vulnérable socialement, et hétérogène tant au niveau du potentiel auditif des personnes sourdes, que de leur parcours éducatif et linguistique (Gillot 1998). Selon leur profil, les personnes sourdes ou malentendantes peuvent utiliser, avec une maîtrise inégale, différents modes de communication reposant sur un canal visuo-gestuel, audio-vocal ou mixte : la Langue des Signes Française (LSF), langue visuelle-gestuelle pratiquée par les personnes sourdes signantes en France ; le français écrit ; ou le français oral dans certaines conditions, seul ou accompagné du Langage Parlé Complété (LPC, code manuel utilisé par une partie des sourds oralisants, pour désambigüiser la lecture labiale).

Pendant les premières années de fonctionnement du centre d'appels, seules sont disponibles les modalités de communication écrite par fax et par SMS. Or, selon le rapport Gillot (1998), 80 % de la population sourde adulte entretient un rapport difficile au français écrit. Ce rapport à l'écrit des personnes sourdes peut se traduire par des productions individuelles qui peuvent mener, pour les agents du 114, à une compréhension approximative ou lacunaire des messages écrits reçus, voire à une incompréhension des situations décrites par les appelants. Or, les objectifs fixés au CNRAU nécessitent avant toute chose que **les agents sourds et entendants comprennent rapidement les messages entrants et formulent des réponses écrites dans un français adapté, c'est-à-dire interprétable par l'appelant sourd.**

Cette situation nous amène à envisager des solutions techniques et scientifiques pour que le CNRAU puisse assurer sa mission. En premier lieu, il s'agit de **caractériser les phénomènes linguistiques et les types d'erreurs propres au français écrit des sourds** afin d'alimenter la formation des opérateurs, mais également pour permettre le développement ultérieur des outils TAL envisagés en soutien à la prise de décision des opérateurs et à la formulation de réponses écrites adaptées. Dans cette perspective, il importera à terme de déterminer l'outillage formel et technique permettant la *correction*, la *complétion*, la *reformulation* et la *traduction* automatiques des messages transmis au centre d'appels.

1.2 Etat de l'art des descriptions du français écrit par des sourds

Le français écrit des sourds est un domaine d'étude peu documenté. Les rares études menées ces 20 dernières années révèlent des phénomènes morpho-syntaxiques et morpho-lexicaux que l'on retrouve dans les productions écrites des sourds québécois (Nadeau 1993, repris dans Nadeau et Machabée, 1998), suisses (Niederberger, 2004) et français métropolitains (Tuller, 2000 et Périni, 2007).

Les personnes sourdes apprennent le français écrit à partir d'un accès au français oral soit lacunaire (restes auditifs, lecture labiale), soit inexistant. La LSF est langue première (L1) dans 10% des cas seulement. Les productions écrites des sourds se caractérisent par une grande variabilité inter-individuelle. Nadeau et Machabée (1998) ont cependant identifié des phénomènes propres au français écrit des sourds qui sont absents des productions des entendants apprenants du français écrit L1 et L2. Ces phénomènes touchent entre autres les rapports localisant/localisé, possesseur/possédé, l'absence de marques morphologiques pour le temps, mais également les redondances lexicales, les confusions de catégories lexicales, et l'utilisation des pronoms - omission illicite des pronoms objets et/ou sujet; confusion entre les 1ère et 3e personnes. Tuller (2000) ajoute l'ellipse de la préposition.

La question suivante se pose tout de même : ces productions constituent-elles une catégorie identifiable (le français écrit des sourds) possédant des caractéristiques communes et des frontières bien délimitées, ou au contraire s'agit-il d'un continuum de productions dont certaines sont plus proches de productions de français écrit d'entendants L1 francophones que d'une autre production de sourd L2. L'écrit des entendants comporte lui-même un certain degré de variabilité inter- et intra- individuel mais ces productions restent dans le champ des « possibles » c'est-à-dire de la grammaire du français écrit. L'étude de notre corpus soutenue par les outils TAL vise également à répondre à cette question.

2 Présentation du corpus FAX-ESSU et de la méthodologie

2.1 Le corpus FAX-ESSU

Quelques extraits choisis : [1] *je ne vais pas lé drocteur* ; [2] *un voiture est brule* ; [3] *elle est tombé son lit* ; [4] *elle tomber tu lit* ; [5] *tombé sur l'escalier* ; [6] *difficile respiratoire*

Le corpus FAX-ESSU (fax Ecrits par des personnes Sourdes en Situation d'Urgence) est un corpus élicité sous forme de fax d'urgence rédigés en français écrit par des personnes sourdes, mises en situation d'urgence. Il a été recueilli dans la perspective d'alimenter la formation des agents du 114. L'élicitation a pris la forme d'un jeu de rôle présenté par une personne sourde en LSF, comportant diverses situations qui nécessitent un appel d'urgence. La consigne donnée aux 17 participants sourds signants et non signants était de rédiger un appel d'urgence manuscrit type fax, sur un temps limité de trente secondes après avoir eu connaissance de la situation critique. 23 situations ont été présentées et ont donné lieu à la production d'énoncés en français. Les productions des situations qui ont posé problème (manque d'entraînement pour la situation 1, erreurs dues à une mauvaise compréhension des situations 7 et 8) ont été éliminées, ainsi que les productions des locuteurs qui comportaient des dessins non traitables à ce stade. Le corpus étudié comporte donc 300 productions correspondant à 15 locuteurs et 20 situations.

Nous présentons ici une première étude lexicale, morphosyntaxique et syntaxique du corpus FAX-ESSU réalisée à la fois manuellement et avec l'aide d'une chaîne de traitement automatique, afin de valider les phénomènes linguistiques décrits dans la

littérature et d'enrichir la connaissance du français écrit sourd.

2.2 Méthodologie

La première étude réalisée sur le corpus FAX-ESSU a pour objet de faire émerger à l'aide d'outils TAL des spécificités du français écrit des sourds relevées ou non dans la littérature. Elle vise également à mettre en évidence les limites d'une chaîne de traitement TAL, dédiée initialement au français standard, dans le traitement du français écrits des sourds.

Pour cela, nous avons travaillé sur trois niveaux : **lexical**, **morphosyntaxique** et **syntaxique**, en procédant à des analyses manuelles et des analyses automatiques corrigées manuellement. Pour le niveau syntaxique (section 3.1), l'analyse du corpus a été réalisée manuellement, en s'appuyant sur le formalisme des Grammaires de Propriétés (Blache, 2000). Ce choix a permis de caractériser les erreurs rencontrées et de les classer par type de contrainte (Propriétés) : **obligation** et **exigence** (omission du noyau syntagmatique et d'autres catégories exigibles), **dépendance** (accord), **précédence** (ordre attendu entre les syntagmes). Une analyse complémentaire a été menée, faisant ressortir des problèmes de sélection de catégorie impliquant régulièrement le niveau lexical, et des phénomènes propres au lexique (section 3.2).

Les analyses automatiques portent sur les niveaux lexical (en cours d'étude) et morpho-syntaxique (section 3.3) ; elles ont été menées avec la chaine de traitement LPLSuite, conçue au Laboratoire Parole et Langage (Blache, Vanrullen, Balfourier, 2006). Cette chaîne a été entraînée sur un corpus de français standard, et évaluée au cours de la campagne nationale d'évaluation des analyseurs syntaxiques EASY 2004-2006, ce qui donne une mesure de sa performance sur le traitement du français écrit standard en comparaison avec les solutions industrielles actuelles. La validation manuelle qui a suivi le traitement automatique a impliqué le travail de trois personnes durant plusieurs mois.

Pour les trois niveaux, l'annotation et la correction manuelles, qui s'appuient sur la segmentation et l'étiquetage automatique du corpus, ont permis de (1) détecter, corriger et comptabiliser les erreurs d'étiquetage morphosyntaxique ; et (2) mettre en évidence différents phénomènes syntaxiques spécifiques au français écrit des sourds, et faire émerger les patrons syntaxiques correspondants.

3 Premiers résultats de l'analyse du corpus FAX-ESSU

3.1 Niveau syntaxique : apport des grammaires formelles

L'analyse syntaxique du corpus repose sur le formalisme des Grammaires de Propriétés (Blache 2000). Cette analyse a permis de mettre en évidence 54 phénomènes syntaxiques dont la fréquence dans les énoncés produits varie de 0,3 % à 62%. Une analyse complémentaire portant sur la sélection de la catégorie morpho-syntaxique a permis de relever en plus 26 phénomènes dont la fréquence va de 0,3 à 16 %.

Dans la figure 1, nous indiquons le risque de rencontrer un type de phénomène donné dans un énoncé, 100% correspondant à une erreur de ce type par énoncé (17 mots en moyenne par énoncé). Sont concernés : **l'omission du noyau syntagmatique**, obligatoire,

(65%) ou d'une autre **catégorie exigée** dans le syntagme (105%) ; les erreurs dans la sélection de la **catégorie** (39%) ou du **lexème** (50%) au sein d'une même catégorie ; une **erreur d'accord** (24%) – genre, nombre, personne – qui concerne un énoncé sur quatre ; et un problème de **précédence** portant sur un mot ou un syntagme mal positionné (8%).

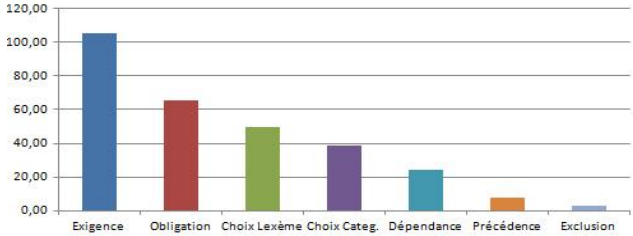

FIGURE 1 – % d'apparition des types de phénomènes rencontrés dans le corpus

Le tableau 1 regroupe nos observations syntaxiques sur les faits les plus remarquables qui apparaissent dans les proportions suivantes (ordre décroissant) : **l'omission du déterminant** (62%), de la **préposition** (35%) et du **verbe** (28%) ; la **sélection de la forme verbale adéquate** (39%) qui concerne pour 28% la catégorie 'verbe' avec un participe passé ou une forme infinitive au lieu d'un verbe fléchi ; et pour 11 % une confusion intra-catégorielle avec un nom à la place d'un verbe fléchi ou d'un participe passé.

Types d'erreurs en % des énoncés	**Omission**		**Accord**	**Sélection de catégorie ou de lexème**
	Obligation	**Exigence**	**Dépendance**	
Verbe	28%	11% (auxiliaire)	5% (genre) 3% (personne) 4% (nombre)	13% (part. passé) 11% (auxiliaire) 4% (infinitif)
Préposition	35%			11% (choix préposition)
Déterminant		62%	5% (genre) 6% (nombre)	3% (err. choix déterminant)
Nom	(1%)	17% (nom ou pronom sujet)		2% (adj) 6% (verbe) 5% (part. passé)
Pronom				8% (err. choix pronom)
Adverbe (nég.)		4% (« ne »)		

TABLEAU 1 –% d'apparition sur le corpus des phénomènes syntaxiques les plus fréquents

Le modèle des Grammaires de Propriétés a permis de rendre compte des phénomènes syntaxiques les plus massifs en termes de fréquence dans les énoncés produits. L'analyse complémentaire a contribué à affiner l'analyse sur la sélection des catégories et des lexèmes, qui ne sont pas prises en compte par les GP.

Cette première analyse fait ressortir des patrons syntaxiques exploitables en TAL. Dans le cadre des GP, ces patrons peuvent prendre la forme d'un assouplissement de certaines contraintes (obligation, exigence, dépendance). A titre d'exemple, un tel assouplissement permettrait de caractériser des **syntagmes verbaux** dont le noyau ou l'auxiliaire seraient absents ou encore dont le noyau serait une catégorie nominale ; de caractériser des

syntagmes prépositionnels sans préposition (!) ; de même pour l'absence du déterminant dans les **syntagmes nominaux** ou encore l'absence du pronom sujet relatif et donc du SN dans les **propositions relatives**.

3.2 Analyse manuelle du lexique : premiers résultats

Les phénomènes qui concernent le **lexique** apparaissent 1,5 fois par énoncé (154%), sans compter les phénomènes de *choix du lexème* que nous avons décrits en section 3.1 (emploi d'un mot pour un autre au sein de la même catégorie). Les erreurs d'accentuation (60%) ne constituent pas seulement un problème lexical, mais peuvent provoquer une confusion de la catégorie au niveau morphosyntaxique. Par exemple, l'absence d'accent sur un participe passé (*blesse/blessé*) qui concerne plus de 15% des énoncés entraîne une mauvaise identification de la catégorie concernée par l'étiqueteur automatique.

En plus de ces ambiguïtés d'origine accentuelle, les problèmes lexicaux tels que les mots inventés (*cacatouches/cacahuètes*) ou ceux orthographiés de façon approximative (*drocteur/docteur*), abrégée ou tronquée, ont également un impact sur la qualité de l'étiquetage automatique, et par suite sur l'analyse syntaxique automatique.

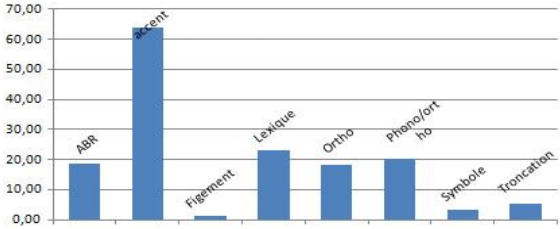

FIGURE 2 – % d'apparition des phénomènes lexicaux (hors choix du lexème)

La situation d'urgence peut être directement mise en cause dans l'usage d'abréviations (*Dr/docteur ; bcp/beaucoup*), de troncations en fin d'énoncé (*doct/docteur* ; *enf/enfant*) ou de symboles (*+/plus*). Ce phénomène peut être comparé aux productions observées dans les SMS. Enfin, les phénomènes concernant les figements ou expressions idiomatiques telles que *perdre les eaux*, *prendre feu*, *tomber dans les pommes*, qui n'ont pas été relevés de manière systématique, devront faire l'objet d'une étude détaillée. Une étude complémentaire du lexique qui s'appuie sur la segmentation automatique est en cours. Elle porte plus particulièrement sur la variabilité lexicale au sein du corpus FAX-ESSU chez les scripteurs sourds.

3.3 Etiquetage morphosyntaxique automatique

Nous avons souhaité vérifier la couverture et la pertinence d'un étiqueteur morpho-syntaxique automatique (POS-tagger) pour l'étude du corpus FAX-ESSU, sachant que le processus d'étiquetage est basé sur l'apprentissage d'un corpus de français écrit standard. Les données quantitatives issues de l'étiquetage et de la segmentation automatiques permettent de dessiner un premier contour très général du corpus. Les 346 phrases reconnues contiennent en tout 4434 mots dont 681 formes distinctes. 181 de ces formes

(soit 26,8%) n'appartiennent pas au lexique du français mais correspondent à des erreurs orthographiques et à des mots inconnus. La figure 3 indique la répartition des catégories présentes dans le corpus sur la base de l'étiquetage automatique (vert) effectué avec l'étiqueteur de la chaîne d'outils LPLSuite. puis corrigé manuellement (rouge).

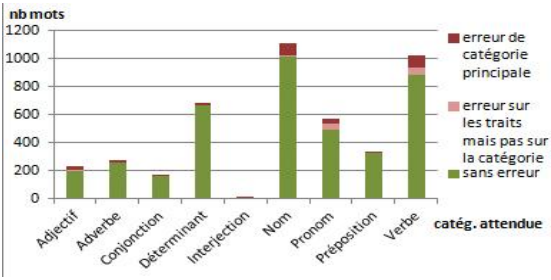

FIGURE 3– Couverture du POS-tagger sur le corpus FAX-ESSU

Une étude plus détaillée des erreurs de catégorisation de l'étiqueteur montre que le nom, le verbe, le pronom et l'adjectif sont moins bien étiquetées que le déterminant, l'adverbe et la préposition. En dehors de l'erreur déjà constatée en français standard sur l'ambigüité adjectif/participe passé, on constate au moins trois types d'erreurs d'étiquetage imputables à la structure syntaxique du français écrit des sourds : verbes interprétés comme noms, pronoms reconnus comme prépositions, noms reconnus comme adjectifs. De fait, le taux important d'omissions et d'erreurs lexicales dans le corpus a un impact significatif sur l'étiquetage. Par exemple, la catégorie Nom est étiquetée 34 fois en tant qu'adjectif , 21 fois comme déterminant, 23 fois en tant que verbe et 13 fois en tant que nom avec cependant des traits erronés (par ex. masc. au lieu de fém.)

4 Discussion des premiers résultats et perspectives

Les phénomènes que nous avons pu relever se recoupent avec les phénomènes pointés par la littérature (cf section 1.2). Les problèmes de compréhension relevés par les lecteurs et annotateurs humains se retrouvent au niveau du traitement automatique sous la forme d'erreurs de catégorisation des lexèmes et des syntagmes. Dès l'étape de l'étiquetage morphosyntaxique, et de façon encore plus proéminente au niveau syntaxique, apparaissent des difficultés et des erreurs de traitement liées aux spécificités du français écrit des sourds. Le TAL offre sur ce sujet un éclairage précis, notamment grâce à un outillage linguistique formel.

Après plusieurs mois de fonctionnement du numéro 114, les messages SMS semblent constituer l'essentiel des appels reçus au centre 114. La constitution d'un corpus SMS d'urgence nécessitera l'anonymisation des données et une gestion des droits et agréments permettant l'exploitation d'un corpus du domaine de la santé (HAS). Dans ces conditions, plusieurs études sont envisagées : dans un premier temps, la méthodologie linguistique et TAL développée et ajustée pour l'étude du corpus FAX-ESSU pourra être appliquée à l'étude des SMS d'urgence. Puis le corpus SMS sourds d'urgence pourra être comparé aux productions SMS sourds métropolitains et de la Réunion récemment recueillies (Blondel *et al.*, 2011), ainsi qu'aux SMS francophones recueillis dans le cadre de la campagne

SMS4Science (Antoniadis *et al.*, 2011). L'étude comparative de ces différents corpus permettra de mettre évidence dans les corpus ESSU ce qui tient à la communication médiée par SMS, ce qui tient au français écrit des sourds et ce qui provient de la situation d'urgence elle-même.

Remerciements

Nous remercions les personnes qui ont contribué à la réalisation de ce travail : Christian, Christine, Joëlle et Roberto pour nourrir nos réflexions sur le français écrit des sourds, les participants sourds au corpus élicité ; Samia et Patricia pour leur travail préparatoire sur le corpus ; Elodie, Fanny et Joy pour leur contribution à l'analyse des données.

Références

ANTONIADIS, G., CHABERT, G., ZAMPA, V. (2011). Alpes4science : Constitution d'un corpus de SMS réels en France métropolitaine. *79eme congrès de l'ACFAS. 9-13 mai 2011, Sherbrook*.

BLACHE, Ph. (2000). Le rôle des contraintes dans les théories linguistiques et leur intérêt pour l'analyse automatique : les Grammaires de Propriétés. *Actes de TALN 2000 (Traitement Automatique des Langues Naturelles)*.

BLACHE Ph., T. VANRULLEN & J.-M. BALFOURIER (2006). Constraint-Based Parsing as an Efficient Solution: Results from the Parsing Evaluation Campaign EASy. *Proceedings of LREC 2006*.

BLONDEL, M., GONAC'H, J., LEDEGEN G. & J. SEELI (2011). Ecriture-sms en Métropole et à La Réunion : « Zones instables et flottantes » du français ordinaire et spécificités du contexte de surdité. Gilles Col (dir.), *Transcrire, Écrire, Formaliser (1). Travaux linguistiques du CERLICO*, 23.

GILLOT, D. (1998). *Le droit des sourds : 115 propositions*. Rapport parlementaire au Premier Ministre.

NADEAU, M. (1993). Peut-on parler de "français sourd"? *Revue de l'Association canadienne de linguistique appliquée* (ACLA), vol. 15, no 2, pages 97-117.

NADEAU, M & D. MACHABEE (1998). Dans quelle mesure les erreurs des sourds sont-elles comparables à celles des entendants ? in Dubuisson C. & Daigle D. (Dir.), *Lecture, écriture et surdité*, Montréal, Logiques édition, pages 169-195.

NIEDERBERGER, N. (2004). *Capacités langagières en langue des signes française et en français chez l'enfant sourd bilingue : quelles relations ?* Thèse de Psychologie, Université de Genève.

PERINI, M. (2007). *La remédiation de l'illettrisme chez les adultes sourds locuteurs de la LSF : travail préparatoire à l'élaboration d'une méthodologie et de supports pédagogiques adaptés.* Mémoire de Master 2, Université Paris 8.

TULLER, L. (2000). Aspects de la morphosyntaxe du français des sourds. *Recherches Linguistiques de Vincennes* 29, pages 143-156, PUV.