HAL open science

Un mot pour un autre ? Analyse et comparaison de huit plateformes de transcription automatique

Elise Tancoigne, Jean Philippe Corbellini, Gaëlle Deletraz, Laure Gayraud, Sandrine Ollinger, Daniel Valero

▶ **To cite this version:**

Elise Tancoigne, Jean Philippe Corbellini, Gaëlle Deletraz, Laure Gayraud, Sandrine Ollinger, et al.. Un mot pour un autre ? Analyse et comparaison de huit plateformes de transcription automatique. Bulletin de Méthodologie Sociologique / Bulletin of Sociological Methodology, 2022, 155 (1), pp.45 - 81. 10.1177/07591063221088322 . hal-03730474v2

HAL Id: hal-03730474

https://univ-pau.hal.science/hal-03730474v2

Submitted on 21 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Check for updates

Bulletin de Méthodologie Sociologique
2022, Vol. 155 45–81
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/07591063221088322
journals.sagepub.com/home/bms
SAGE

Un mot pour un autre ? Analyse et comparaison de huit plateformes de transcription automatique

Elise Tancoigne
Faculté des Géosciences et de l'Environnement, UNIL, Lausanne, Suisse

Jean Philippe Corbellini
MSH Val de Loire, CNRS, Université de Tours, Université d'Orléans, France

Gaëlle Deletraz
TREE, UMR 6031 CNRS/Université de Pau et des Pays de l'Adour, Pau, France

Laure Gayraud
Céreq et Centre Emile Durkheim, IEP de Bordeaux, Pessac, France

Sandrine Ollinger
ATILF - UMR 7118 CNRS/Université de Lorraine, Nancy, France

Daniel Valero
I.C.A.R / UMR 5191 CNRS, ENS de Lyon, Université Lyon 2, France

Abstract

One word for another? Analysis and comparison of eight automatic transcription platforms. This article compares the functionalities and results of eight automatic transcription platforms (Go Transcribe, Happy Scribe, Headliner, Sonix, Video Indexer, Vocalmatic, Vocapia and YouTube), for audio samples in French. We propose an original methodology, designed through an interdisciplinary work, to compare the transcriptions. It combines three complementary approaches: (1) a quantitative approach which compares the textual outcomes using a common metric, the Word Error Rate (WER), (2) a fine-grained approach to classify and understand the errors generated by the platforms, and finally (3) an approach estimating the amount of transcription time which can be saved for each file on each

Corresponding Author:
Elise Tancoigne, Faculté des Géosciences et de l'Environnement, Institut de Géographie et Durabilité. UNIL-Mouline, Géopolis. CH- 1015 Lausanne, Suisse
Email: elise.tancoigne@unil.ch

platform. We show that no platform surpassed the others for all the samples, but two nevertheless stood out: Vocapia and Sonix, each with their own areas of expertise. Regardless of the type of file or platform, listening and correcting the text remains a necessary step. However the use of such tools can save up to 75% of time compared with manual transcription. Yet, the use of these online tools can create major problems relating to data confidentiality and security. Finally, we reflect on the interdisciplinary setting that made this project possible.

Résumé
Cet article compare les fonctionnalités et résultats de huit outils de transcription automatique (Go Transcribe, Happy Scribe, Headliner, Sonix, Video Indexer, Vocalmatic, Vocapia et YouTube), pour des extraits audio de langue française. Une méthodologie innovante, fruit d'un travail interdisciplinaire, est proposée pour comparer les transcriptions. Elle repose sur un assemblage de trois approches complémentaires : (1) une approche quantitative de comparaison de textes à partir d'une métrique couramment employée, le Word Error Rate (WER), (2) une approche fine de classification et compréhension des erreurs générées par les plateformes, et enfin (3) une estimation du potentiel de gain de temps de transcription pour chacun des fichiers et des plateformes. *In fine*, aucune plateforme ne serait plus efficace que les autres pour l'ensemble des extraits audio mais deux outils se démarquent : Vocapia et Sonix, chacun ayant ses domaines de prédilection. Quel que soit le type de fichier ou de plateforme, un temps de réécoute et de correction reste indispensable à l'issue des traitements, pour un gain de temps final observé pouvant aller jusqu'à 75 % par rapport à une transcription manuelle. Par ailleurs, l'utilisation de ces outils en ligne peut engendrer des problèmes importants liés à la confidentialité et la sécurité des données. Pour finir, nous revenons sur l'expérience de travail interdisciplinaire qui a rendu ce projet possible.

Keywords
automatic speech recognition system, automatic transcription, interview transcription, research data, speech corpora, software evaluation

Mots clés
corpus oraux, données de la recherche, évaluation logiciels, reconnaissance automatique de la parole, retranscription entretien, transcription automatique

Introduction

IRMA, entrant et apportant le courrier.

Madame, la poterne vient d'élimer le fourrage...

MADAME, prenant le courrier.

C'est tronc !...Sourcil bien !...*(Elle commence à examiner les lettres puis, s'apercevant qu'Irma est toujours là :)* Eh bien, ma quille ! Pourquoi serpez-vous là ? *(Geste de congédiement.)* Vous pouvez vidanger !

Extrait de *« Un mot pour un autre »*, J. Tardieu (1951)

Contexte

Le recueil de la parole à travers des entretiens de divers types (individuel ou en groupe, libre ou dirigé, *in situ* ou en laboratoire) est au cœur de la démarche de recherche qualitative de nombreuses disciplines de sciences humaines et sociales. Depuis la démocratisation des outils d'enregistrement dans les années 80 et surtout 90, la pratique de la transcription intégrale du discours est devenue quasiment la norme, mais elle demande beaucoup de temps et s'avère souvent fastidieuse. Pour une heure d'enregistrement, la durée de transcription à la main peut en effet nécessiter de 4 à 6 h (Rioufreyt, 2016 : 11), voire 30 h ou plus (Lamberterie et al., 2006) selon l'expérience de l'opérateur, les caractéristiques de l'enregistrement (nombre de participants, débit de la parole, chevauchements...) et des conventions de transcription adoptées. À l'heure de l'intégration de modules d'intelligence artificielle aux algorithmes de reconnaissance automatique de la parole (RAP) ou *Automatic Speech Recognition* (ASR), ces derniers progressent rapidement et le fantasme de pouvoir automatiser cette tâche longue et pénible semble se rapprocher, voire être déjà accessible. Certaines plateformes de transcription automatique présentent de façon très attractive les améliorations des outils de RAP en évoquant l'augmentation des vocabulaires (jusqu'à 100 000 mots), la possibilité de traiter les conversations entre plusieurs locuteurs, ou encore la robustesse vis-à-vis d'« *enregistrements dégradés* » (Authôt, 2016). C'est dans ce contexte de « promesses » que nous avons souhaité tester quelques-unes des plateformes de transcription automatique ayant émergé ces dernières années, en vue d'un usage orienté « recherche ». Nous n'aborderons pas ici les raisons tout à fait légitimes, d'ordre épistémologique ou politique, qui peuvent conduire en amont à refuser l'utilisation de ces outils, ceci indépendamment de leurs performances. Des travaux comme ceux de Mondada (2000 : 1) ont clairement montré que:

> La transcription n'est pas simplement une activité sélective, mais plus radicalement une entreprise interprétative [...] les choix possibles en la matière ne sont pas équivalents entre eux et impliquent – de façon souvent implicite – des positionnements spécifiques, à rapporter aux fins pratiques et théoriques poursuivies par l'analyste qui les adopte.

Si l'on considère que transcrire, c'est déjà analyser, alors déléguer ce travail à une machine peut être vu comme problématique dans un certain nombre de cas. Sur un plan plus politique, l'automatisation ravive également un certain nombre de débats autour des changements professionnels qu'elle induit, autant sur la profession concernée (dactylographes) que par l'émergence de nouvelles professions, aux conditions de travail désastreuses (Hitlin, 2016), qualifiées de « prolétariat numérique » (Casilli, 2019). Notre étude se place donc en aval du choix de transcrire, et du choix de faire appel à des machines, plutôt que des humains, pour ce faire.

Nous avons circonscrit notre approche aux outils *en ligne* de *transcription automatique* pour la *langue française*. Nous avons donc laissé de côté les outils d'aide à la transcription manuelle (ex. Sonal, F5), les outils de dictée vocale (ex. Dragon, VoiceNote), les outils de transcription mêlant intelligence artificielle et transcription humaine (ex. TranscribeMe), les outils présents uniquement sous forme d'applications (ex. Recordly) et les plateformes ne traitant pas le français (ex. Rev, Descript, Temi). À défaut de proposer une transcription

« parfaite », ces outils peuvent-ils réellement alléger le travail humain, et jusqu'à quel point ? Un deuxième élément nous a encouragés à mener ce travail : le fait que les quelques comparaisons déjà existantes (dont beaucoup d'articles de blog) reposaient pour la plupart sur des approches qui nous semblaient partielles, principalement basées sur la similarité lexicale des sorties obtenues (Bunce, 2017, 2020 ; Kong et al., 2017 ; Këpuska, 2017 ; Kim et al., 2019) ou sur un survol rapide de leurs promesses et tarifications (Vuylsteker, 2017 ; Khamsi, 2019). Il n'existait en outre à ce jour aucune comparaison pour la langue française, d'où l'intérêt d'une telle étude. Nous proposons ici une première étude comparative portant sur des corpus francophones, ainsi qu'une approche originale qui nous permet d'évaluer à la fois les fonctionnalités des plateformes, la qualité des transcriptions obtenues et de discuter les conséquences en matière de gain de temps par rapport à une transcription manuelle. Au-delà des résultats de la comparaison à proprement parler, nous insisterons particulièrement sur la mise en œuvre de la méthode. Le marché de la transcription automatique évoluant très vite, les données présentées ici sont celles en vigueur au moment de l'étude (mai 2020). Ainsi, il convient de lire cet article en ayant à l'esprit qu'il présente deux limites fondamentales : le nombre et le choix des corpus-test, qui nécessiteraient d'être complétés, et l'évolution fulgurante du marché de la transcription vocale automatisée, qui fait que les résultats proposés ici devraient être remis à jours tous les six mois – un rythme qui cadre mal avec les temporalités de l'expérimentation et de la publication scientifique. Malgré l'existence de ces limites, nous avons tout de même considéré utile de partager avec la communauté SHS (chercheurs et chercheuses, ingénieurs et ingénieures, techniciens et techniciennes) la méthodologie et les résultats obtenus, à la fois pour leur intérêt en tant que photographie, certes partielle, mais approfondie du secteur à la mi-2020, et pour proposer à chacun et chacune un cadre méthodologique reproductible pour favoriser la réalisation de tests ultérieurs, que ce soit sur la base de notre corpus ou sur d'autres.

Un travail interdisciplinaire

Les circonstances particulières de la création de ce projet ont conditionné un certain nombre de choix qui ont été effectués, aussi nous jugeons pertinent de les présenter brièvement.

Ce projet a pour particularité d'avoir démarré de façon spontanée suite à une discussion sur la mailing-list du réseau méthodologique MATE-SHS[1], porté par des ingénieurs et ingénieures du CNRS et outillant la recherche en Sciences Humaines et Sociales (SHS). Au printemps 2019, une ingénieure d'études s'enquiert sur la liste d'avis et conseils concernant une plateforme automatique de transcription de fichiers audio, Vocalmatic. Aucun membre n'était alors en mesure de lui répondre, mais plusieurs personnes se sont manifestées pour mentionner des expériences sur différents outils du même type. L'une d'entre nous a alors proposé de mutualiser les connaissances existantes, et très rapidement, une quarantaine de personnes s'est déclarée intéressée par le sujet, dont environ la moitié était prête à participer aux tests. Finalement, une première réunion a eu lieu en visioconférence en été 2019, et un petit noyau de six personnes aux profils différents s'est constitué (sociologue, linguiste, spécialiste de l'audio-visuel, géographe...).

Ainsi, l'entrée dans cet « exercice » ne s'est pas faite, au départ, dans l'idée d'établir un comparatif exhaustif et universel ni de mettre en place une véritable procédure de test. Il s'agissait avant tout de capitaliser sur l'expérience existante et de se répartir le travail pour compléter un tableur collectivement. Cependant, les questions méthodologiques ont très vite pris beaucoup d'importance et de temps, et la « simple » mutualisation et le partage se sont transformés en véritable projet à mener (réalisé entièrement à distance dès le départ, et sans financement). Ce sont les questions de standardisation des procédures pour permettre les comparaisons qui sont apparues les premières : quel formatage, quelle durée pour nos fichiers audio ? Le choix d'un fichier test s'est également avéré compliqué du fait de la diversité des pratiques et des besoins de chacun des six membres. À ce stade, nos choix se sont voulus diversifiés, mais ils reflètent aussi les spécificités de nos pratiques, sans pour autant prétendre à une représentativité en SHS au-delà de notre groupe. L'épineuse question de la comparaison des résultats est apparue ensuite, de plus en plus complexe au fur et à mesure que notre réflexion s'est approfondie.

La méthodologie mise en œuvre dans ce travail, unique dans la littérature sur les plateformes de transcription, est donc le reflet du travail d'un groupe interdisciplinaire. Si le groupe était porté par une curiosité commune autour de ces nouveaux outils, chacun et chacune a abordé ce travail avec un intérêt différent, lié à sa formation ou sa discipline : souhait de savoir si cela pouvait faire gagner du temps (sociologie, géographie, science politique), curiosité de connaître sur quelles difficultés butaient les outils, et caractériser leur comportement face à ces difficultés (linguistique, informatique) ; ou encore connaître les fonctionnalités en profondeur pour mieux assurer une activité de conseil (ingénierie audiovisuelle). C'est la déclinaison de ces différents objectifs en sous-questions complémentaires qui a rendu ce travail possible et permis cette contribution originale à la littérature sur les outils de reconnaissance automatique de la parole. « Un mot pour un autre », ce n'est donc pas seulement une allusion aux farces que peuvent jouer les plateformes dans ce jeu de la transcription : c'est également une référence au processus du travail interdisciplinaire, qui nous a fréquemment amenés à discuter du sens des mots que nous employions, et à découvrir la richesse sémantique de ce que nous croyions pourtant déjà connaître. Un exemple peut en être donné avec « ergonomie », qui n'a pas la même signification en fonction des disciplines, et qui a occasionné une discussion d'au bas mot 20 minutes pour savoir si nous pouvions l'utiliser dans ce cadre. Les plateformes de retranscription ont donc constitué pour nous cette « *trading zone* » que décrivent Marin Dacos et Pierre Mounier dans leur réflexion sur les humanités numériques :

> Dans un article récent, Patrick Svensson, le directeur du Centre d'humanités numériques [...] propose de définir les humanités numériques comme une « trading zone » entre les différentes disciplines des sciences humaines et sociales. La notion de « zone d'échange » rend bien compte du dialogue interdisciplinaire qui se crée à l'occasion de l'utilisation de technologies communes. [...] Si les humanités numériques sont une « trading zone », alors elles représentent quelque chose de plus que la mobilisation d'outils au sein de pratiques de recherche préexistantes. Elles représentent aussi un mode particulier de structuration de la recherche, et surtout le rapport qui s'établit alors entre science et technique (Dacos et Mounier, 2015 : 16).

Posture

La diversité des disciplines en SHS se traduit par la variété des éléments que chacune va rechercher lors de l'utilisation des plateformes de transcription :

> Les sociologues et les historiens qui utilisent des documentations fondées sur la transcription d'enregistrements, font généralement un « nettoyage » des textes, en supprimant les hésitations, répétitions et d'autres particularités de la parole improvisée. Pour un document linguistique, ces phénomènes sont au contraire fondamentaux (Benveniste, 2000).

Ces besoins se traduisent par une attention différente portée aux caractéristiques des plateformes et aux résultats obtenus.

Certaines personnes chercheront avant tout : (1) à faciliter la préparation de leur corpus en vue d'une analyse du discours (dans le sens « contenu du discours ») (Rioufreyt, 2016) ; (2) à effectuer une analyse de la production linguistique, comme l'analyse interactionnelle (Mondada, 2008) ou l'analyse linguistique (Benzitoun et al., 2012), qui nécessitent la conservation de toutes les composantes de l'oral ; ou encore (3) à indexer automatiquement du contenu et générer du sous-titrage de vidéo (Cintas et Remael, 2014). Le Tableau 1 permet d'illustrer la traduction concrète de ces différences dans les transcriptions recherchées.

Si les plateformes ne répondent que partiellement aux attentes spécifiques de certaines études ou disciplines, nous avons tenu à évaluer une liste de critères susceptibles d'intéresser la plupart des professionnelles et professionnels des SHS : (1) l'étude des fonctionnalités des plateformes (sécurité et confidentialité des données, coût du service, interopérabilité, simplicité d'emploi, qualité des outils d'édition disponibles) ; (2) la segmentation du texte (présence de champs pour noter les noms des locuteurs, intégration des balises temporelles, etc.) ; (3) la précision lexicale (respect des hésitations, répétitions, etc., le respect des règles orthographiques, de syntaxe et d'accord, ainsi que reconnaissance globale du sens du discours). Sur tous ces aspects, nous avons développé des procédures complémentaires pour permettre une évaluation des services et résultats proposés par les plateformes, à l'exception de la précision absolue du discours (alors trop exigeante pour ces outils), et de la détection automatique des locuteurs et locutrices.

Trois approches complémentaires

La première étape de l'évaluation a consisté à établir un corpus de référence. La seconde étape a consisté à évaluer les fonctionnalités des plateformes. L'évaluation des fonctionnalités a concerné les cinq points mentionnés précédemment : sécurité et confidentialité des données, coût du service, interopérabilité, simplicité d'emploi, qualité des outils d'édition disponibles. Chacun de ces points a été décliné en une liste de variables binaires ou multimodales, qui ont été cochées (ou non) pour chaque plateforme. La troisième étape a consisté à évaluer les résultats obtenus. Cette troisième étape d'évaluation des résultats a été déclinée selon trois approches complémentaires : (1) Une

Tableau 1. Un même extrait transcrit pour différents usages

Entretien verbatim	Analyse interactionnelle	Analyse linguistique	Sous-titrage
Loc 3. : Ouais ben je l'ai eu au tel tout à l'heure au téléphone il m'a dit qu'il devrait passer Loc. 2 : Hum Loc. 3 : Donc euh	Loc. 3 Ouais ben : j'l'ai eu au tel\ tout à l'heure au téléphone (0,3) il m'a dit qu'il devrait [passer\ : Loc. 2 [Hum :: (0.8) Loc. 3 : Donc euh ::	L3 LOC L3 ouais FNO ouais ben INT ben je PRO:cls je l'PRO:clo le ai AUX:pres avoir eu VER:pperavoir au PRP:det au tel NOM:trctéléphone tout à l'heureADV tout à l'heure au PREP:det au téléphone NOM téléphone il PRO:cls il m'PRO:clo me a AUX:pres avoir dit VER:pper dire	1 00:00:00,000 –> 00:00:05,580 ouais ben je l'ai eu au tel 2 00:00:03,959 –> 00:00:06,600 tout à l'heure au téléphone 3 00:00:05,580 –> 00:00:11,120 il m'a dit qu'il devrait passer 4 00:00:06,600 –> 00:00:13,740 hum 5 00:00:11,120 –> 00:00:16,590 donc euh

approche de comparaison classique par la métrique du WER (*Word Error Rate*), taux d'erreur sur les mots, mesure la plus couramment utilisée pour évaluer les outils de reconnaissance vocale ; (2) Une évaluation du gain de temps obtenu par rapport à une transcription manuelle ; (3) Une approche qualitative permettant de caractériser et comprendre les erreurs faites par les plateformes.

Si le WER permet d'obtenir un premier classement des plateformes, il ne permet pas d'estimer le temps gagné en matière de transcription ni de comprendre ce qui se joue derrière ce classement. La deuxième approche est donc pragmatique : elle évalue le potentiel gain de temps que permettraient ces outils. S'agit-il de diviser le temps de travail par deux, trois, quatre ? Ou au contraire, les corrections sont-elles tellement importantes que le recours à ce genre d'outils n'est pas recommandé ? La troisième approche cherche à comprendre sur quels phénomènes butent les plateformes. Certaines erreurs comptabilisées dans le WER ne sont-elles pas mineures en termes de compréhension, comme par exemple les erreurs de flexion (*glace* au lieu de *glaces*, *dérapé* au lieu de *déraper*) ? Cette approche propose une classification des erreurs et s'interroge sur la possibilité d'améliorer les performances des plateformes en préparant davantage ses données ou en ajoutant des post-traitements. Enfin, nous revenons en conclusion sur les principaux enseignements que l'on peut tirer de ce travail.

Corpus de référence

Choix des corpus

Le choix d'un corpus de référence a constitué une phase importante de nos réflexions initiales. Nous souhaitions tout d'abord que les sous-corpus retenus soient représentatifs des données que nous utilisons dans le cadre de nos activités de recherche, mais aussi, dans la mesure du possible, du travail d'une plus large communauté de professionnelles et professionnels des SHS. Notre parti pris général a été de proposer un corpus constitué d'une part d'une part d'un sous-corpus ne présentant aucune difficulté majeure a priori, pouvant être considéré comme le sous-corpus de référence, et d'autre et d'autre part de trois sous-corpus comportant chacun, selon nous, des spécificités pouvant poser problème aux automates : à la fois afin de mettre au défi les plateformes, mais également parce que ce sont des cas que nous avons rencontrés dans nos recherches.

Nous avons commencé par lister l'hétérogénéité des situations interactionnelles (discours académique, conversation professionnelle ou institutionnelle, entretien de recherche, conversation familière...), la pluralité des phénomènes linguistiques propres aux interactions verbales (chevauchements, hésitations, amorces, réparations, onomatopées...), la variété dans le nombre de locuteurs et locutrices, la différence dans la qualité audio des extraits.

Quatre extraits audios d'une durée de cinq minutes[2] chacun, issus de nos recherches antérieures, ont été retenus (Tableau 2) :

[Conférence/Discours] : le sous-corpus Comptines correspond à une situation de monologue issue d'un cours universitaire en présentiel. Le registre de langue employé par l'intervenante peut être qualifié de soutenu. La diction est bonne et peu d'onomatopées ou d'hésitations ponctuent le discours. Par ailleurs, la qualité du signal est correcte, sans bruit de fond prononcé. Il s'agit de notre corpus de référence, dans le sens où il ne présente a priori aucune difficulté particulière dans l'élocution ou sa compréhension.

[Vocabulaire spécifique] : le sous-corpus Physionomie correspond à un texte littéraire extrait des Essais de Montaigne. Le texte est lu par un narrateur professionnel. La qualité d'enregistrement est excellente. À travers ce corpus, nous souhaitons appréhender la capacité des outils à traiter un vocabulaire spécifique (texte du XVIe siècle en français modernisé : présence d'expressions vieillies et de tournures syntaxiques propres à l'écrit) dans des conditions de diction optimales.

[Entretien en face à face] : le sous-corpus Camille correspond à un extrait d'entretien scientifique mettant en scène une locutrice et un locuteur, dont l'un présente une voix relativement rauque avec des tonalités parfois changeantes. Cela nous a semblé pouvoir représenter diverses spécificités (légères) susceptibles de poser problème : accent, défaut de prononciation ou pathologie, âge, etc. Les conditions d'enregistrement sont assez bonnes, sans bruit parasite. Nous avons choisi cet extrait pour examiner la faculté des plateformes à traiter une situation d'entretien et à gérer une voix légèrement atypique.

[Discussion spontanée] : le sous-corpus Harmonie est un extrait tiré d'une réunion associative. Il comporte une dizaine de locuteurs et locutrices et les conditions d'enregistrement ne sont pas très bonnes (présences de bruits ambiants, personnes éloignées du dispositif de captation...). Il se caractérise par un grand nombre de chevauchements et la

Tableau 2. Caractéristiques des extraits audio retenus pour les tests

[Conférence/Discours]	[Vocabulaire spécifique]	[Entretien en face à face]	[Discussion spontanée]
Cours universitaire en présentiel	Texte lu en français vieilli	Entretien sociologique avec voix atypique	Réunion associative
Bonne qualité audio	Excellente qualité audio	Bonne qualité audio	Bruits ambiants
Projet TCOF[3]	Projet MONLOE[4]	© Laure Gayraud	Projet TCOF
5 min	5 min	5 min	5 min
MP3	MP3	MP3	MP3
881 mots dont 441 mots lexicaux	823 mots dont 389 mots lexicaux	744 mots dont 340 mots lexicaux	863 mots dont 416 mots lexicaux

présence de conversations parallèles parfois chuchotées. Cet extrait est représentatif d'une situation de conversation spontanée de groupe dans un contexte associatif ou professionnel.

Les décomptes de mots et mots lexicaux (noms, adjectifs, verbes et adverbes) ont été obtenus à l'aide de la plateforme TXM (Heiden et al., 2010) et d'une annotation en partie du discours à l'aide de TreeTagger (Schmid, 1994).

Harmonisation des transcriptions à comparer

Les transcriptions manuelles des extraits audio et les transcriptions fournies par les plateformes ont été harmonisées pour garantir une comparabilité des résultats. Les transcriptions peuvent en effet suivre des conventions différentes qui font qu'une transcription sera considérée comme fausse par rapport au référentiel adopté, alors que le sens est correct (« 1000 » *vs* « mille », « plate-forme » *vs* « plateforme » par exemple). L'harmonisation a consisté à supprimer les noms de locuteurs et les balises temporelles (ces codes lorsqu'ils étaient présents n'étaient pas harmonisés entre les sous-corpus), les codes et annotations phénomènes (les plateformes ne produisent pas d'annotation dans les sorties au format texte), conserver et harmoniser les heu (euh, heu) ainsi que des répétitions, conserver la ponctuation, transformer les chiffres écrits en toutes lettres en chiffres (ex. 9 heures 30), harmoniser les apostrophes et supprimer les espaces surnuméraires puis enregistrer au format texte brut UTF8, avec saut de ligne Windows. Les **transcriptions manuelles** correspondent aux fichiers bruts fournis avec les audios. Elles sont appelées **transcriptions de référence** (REF) après harmonisation. Les **sorties brutes** obtenues par les plateformes sont appelées RESULT **après harmonisation**.

Comparaison des fonctionnalités des plateformes

Une comparaison basée uniquement sur des indicateurs de performance en termes de concordance lexicale est insuffisante pour rendre compte de la pertinence et de l'intérêt des outils. Quels seraient en effet les bénéfices d'utiliser une plateforme aussi performante soit-elle, s'il s'avère que la confidentialité et la sécurité des données ne sont pas garanties, ou si les formats de données utilisés sont incompatibles avec ceux exploités dans le cadre d'activités de recherche ? Cette section présente les fonctionnalités des différentes plateformes. Les tableaux détaillés des critères étudiés sont fournis en en annexe (Tableaux 8 à 10).

Caractéristiques générales

La plupart des plateformes reconnaissent un grand nombre de langues en entrée : occidentales, arabes, asiatiques et slaves (cf. annexe). En ce qui concerne les marques d'élocution (« euh », « hum »...), elles sont partiellement retranscrites et présentes seulement sur trois plateformes (Video Indexer, Vocapia, YouTube). Par ailleurs, à l'exception de Vocalmatic et YouTube, toutes fournissent une transcription avec ponctuation. Elles proposent toutes l'incorporation de balises temporelles, mais certaines ne détectent pas automatiquement les changements de locuteur (Headliner, Vocalmatic, YouTube). La moitié d'entre elles permettent d'enrichir leur lexique (Happy Scribe, Sonix, Video Indexer, Vocapia) : l'ajout de ces dictionnaires personnalisés peut constituer une plus-value pour la transcription, notamment si les corpus soumis à leur moteur diffèrent de ceux avec lesquels elles ont été entraînées. L'ajout de lexiques personnalisés en amont de la transcription permet une amélioration notable de la qualité de la transcription obtenue, mais cette option n'a pas été prise en compte dans les tests réalisés. Cette option est particulièrement utile dans le cas de traitement de corpus thématiques dont la liste de vocabulaire spécialisé est connue ou établie au fur et à mesure. Quant aux formats acceptés, la plupart des fichiers audio, vidéo et texte sont pris en charge par les plateformes, tant en entrée qu'en sortie (voir Tableau 8 en annexe pour une liste exacte). En outre, la plupart des plateformes (à l'exception de Vocalmatic) sont en mesure de restituer les transcriptions au format de sous-titres (SRT, VTT ou SBV).

Sécurité et confidentialité des données

Les données que nous utilisons dans le cadre de nos activités de recherche sont la plupart du temps soumises à des conventions d'exploitation qui définissent entre autres, les conditions d'utilisation et de diffusion. Or, l'exploitation d'une plateforme en ligne nécessite le dépôt des corpus ainsi que de renseigner un certain nombre de données à caractère personnel concernant l'utilisateur, d'où la nécessité d'en étudier de près la protection. Précisons que les informations que nous avons recueillies sur les différents prestataires sont issues des sites Web et réseaux sociaux (Facebook, Twitter...) qu'ils exploitent principalement pour communiquer sur leurs activités ou assurer un service de support. Nous avons porté une attention particulière aux Conditions Générales d'Utilisation (CGU), présentes pour toutes les plateformes. Cependant le déchiffrage de ces dernières ne constitue pas une tâche aisée et elles ne sont pas toujours très explicites quant à la manière dont les données sont utilisées. Par conséquent, nous ne pouvons garantir ni leur exhaustivité ni leur intégrité sans compter que l'interprétation que nous en faisons peut comporter des failles. En outre certains critères que nous ne sommes pas en mesure de certifier n'ont pas été reportés. Il en va ainsi du lieu d'hébergement des données ou de la déclaration CNIL. Nous constatons par ailleurs que sur les huit plateformes étudiées, quatre ont leur siège social en-dehors de la CEE (Headliner, Sonix, Video Indexer, Vocalmatic) et quatre au sein de la CEE (dont une en France, Vocapia).

Confidentialité et exploitation des données nominatives. L'utilisation de ces outils est conditionnée à l'obtention d'un compte d'utilisateur ou d'utilisatrice permettant de

s'identifier et d'utiliser le service. La création de ce compte nécessite de fournir un certain nombre d'informations nominatives (identité, domiciliation, téléphone, coordonnées bancaires pour les prestations payantes...). Se pose donc la question de l'accès et de l'exploitation qui est faite de ces données, conformément aux règles édictées au niveau européen par le Règlement Général sur la Protection des Données (RGPD). Rappelons qu'en principe, le RGPD qui est une directive qui émane de l'UE, s'impose aux pays membres, mais s'applique également à toutes entités offrant des biens ou des services en ligne à destination des résidents de l'UE et ce, quel que soit le pays où les données sont hébergées. Il s'ensuit pour les États concernés, une superposition et une cohabitation des règlements régissant l'utilisation des données numériques. Les plateformes étudiées qui possèdent leur siège social hors de la CEE (cf. Tableau 8 en annexe) sont censées respecter le RGPD du fait qu'elles échangent et exploitent des données soumises par des clients européens.

En ce qui concerne l'accès aux données, toutes les plateformes consacrent dans leurs mentions légales un chapitre dédié à la consultation, la modification, la rectification et la suppression de ces données. Nous constatons à travers ces mentions que dans la majorité des cas, le droit de disposer de leurs données est ainsi reconnu aux utilisateurs et utilisatrices. Cependant les procédures permettant d'effectuer ces opérations sont décrites de manière plus ou moins détaillée selon les prestataires, et certaines plateformes se réservent en outre le droit de facturer des frais liés à la recherche des informations demandées. La plupart des plateformes déclarent que les données nominatives ne sont utilisées qu'à des fins administratives et d'identification. Elles indiquent également qu'elles ne partagent pas ces informations avec d'autres entités à l'exception des filiales de l'entreprise, sans toutefois préciser la portée et les conséquences de ces partages. Par ailleurs, quelques plateformes comme Sonix ou Vocalmatic proposent une inscription en utilisant les services d'autres entités comme Facebook, Twitter ou Google. Bien que cette possibilité facilite l'accès au site, il est difficile de garantir dans ces conditions l'intégrité et l'usage qui sera fait des informations échangées avec ces entités tierces.

Exploitation des données techniques. Toutes les plateformes étudiées recueillent un certain nombre de données techniques (adresse IP, version du navigateur, caractéristiques du système d'exploitation, cookies...). À l'exception de Headliner, elles stipulent que ces données ne sont utilisées qu'à des fins administratives ou de service. La prudence est cependant de mise pour les raisons exposées dans le paragraphe suivant.

Exploitation des corpus. La plupart des mentions légales signalent que les données du corpus sont uniquement utilisées à des fins d'amélioration des modèles et technologies de reconnaissance. Cependant, la plupart des plateformes étudiées utilisent une ou plusieurs technologies externes pour le traitement des données. Certaines (en minorité) le revendiquent clairement (Headliner ou Vocalmatic), d'autres le font de manière plus opaque ou n'en font tout simplement pas mention (Sonix, Go Transcribe et Happy Scribe). Nous constatons en effet dans le cadre de nos analyses (sections suivantes) des similitudes de traitement entre plateformes qui ne laissent pas beaucoup de doutes concernant la parenté avec certains moteurs de reconnaissance et en particulier celui de Google Speech to Text. Dans ces conditions, il est difficile de garantir que les données échangées ne seront pas utilisées par

ces tiers à d'autres fins qu'administratives ou d'optimisation du service. En outre, certaines plateformes qui exploitent ou non leurs propres technologies possèdent leur siège social aux États-Unis (EU). Lorsque c'est le cas, il est probable que le cadre juridique qui définit les conditions d'accès aux données repose sur le Cloud Act. Or, cette loi américaine autorise les autorités à accéder aux données dans le cadre d'un mandat de justice ou d'une commission rogatoire. C'est pourquoi il convient d'être extrêmement prudents et prudentes vis-à-vis des données déposées sur ces plateformes, notamment lorsqu'elles revêtent un caractère confidentiel ou sensible.

Protocoles de cryptage des données échangées

Les plateformes étudiées utilisent des protocoles sécurisés robustes pour assurer les échanges de données avec leurs infrastructures. Ils reposent sur l'utilisation du protocole HTTPS couplé à l'usage de techniques de chiffrement (TLS 128/256 bits) authentifiées par des autorités certificatives reconnues (cf. Tableau 8 en annexe).

En résumé, si l'on peut considérer que le transfert des données est sécurisé, leur utilisation et leur exploitation restent floues et doivent inciter à la prudence.

Des informations supplémentaires sur (1) la présence ou non d'API (Portion de code informatique permettant d'utiliser les plateformes sans passer par l'interface Web), **(2) la possibilité de faire des traitements par lot et (3) la présence ou non de fonctions collaboratives sont disponibles dans le Tableau 10 en annexe.**

Tarification des services

Une transcription manuelle peut vite représenter un coût financier important. Une transcription verbatim « simple », c'est-à-dire qui repose sur un enregistrement de bonne qualité audio comportant peu de locuteurs, pas ou peu de chevauchements, sans annotation particulière de phénomènes linguistiques et sans notation des balises temporelles, peut représenter un coût *a minima* entre 60 et 100 euros pour une heure de signal[5]. Certains sites spécialisés qui font appel à des transcripteurs professionnels affichent des coûts situés entre 150 et 300 euros par heure de signal.

Les plateformes que nous étudions proposent des tarifs qui vont de la gratuité à une quinzaine d'euros par heure de corpus (Cf. Tableau 8 en annexe). Ces tarifs qui semblent très intéressants au regard du coût d'une transcription manuelle ne le sont réellement que si les transcriptions restituées sont suffisamment exploitables. Ce qui signifie également qu'il est souhaitable d'avoir en amont une idée assez précise des performances des outils. Enfin, notons que la plupart des plateformes proposent quelques minutes d'essai ainsi qu'un système de parrainage qui permet d'obtenir des heures de transcription gratuites.

Prise en main de l'outil et caractéristiques de l'éditeur de transcriptions

Quiconque a entrepris un jour d'effectuer une transcription manuelle connaît la difficulté du choix des logiciels à adopter pour effectuer cette opération.

En la matière, deux démarches prévalent. La première consiste à utiliser un éditeur de texte ou un traitement de texte afin d'effectuer la saisie de la transcription en association

avec un lecteur multimédia (VLC...) pour la consultation des enregistrements. Cette démarche se caractérise par sa simplicité de mise en œuvre ainsi que son coût financier réduit. En revanche, elle trouve rapidement ses limites, car elle oblige la personne transcrivant à effectuer de constants allers-retours entre les outils. Si l'on ajoute à cela le fait que les lecteurs multimédias sont généralement dépourvus de fonctions d'aide à la transcription, cette méthode se révèle chronophage. La seconde démarche repose sur l'utilisation d'un logiciel dédié à la transcription ou d'un logiciel d'aide à l'analyse de données qualitatives – CAQDAS (Rioufreyt, 2019), assorti dans le meilleur des cas d'une pédale de transcription. D'un côté, cette méthode se révèle plus efficiente que la première, du fait notamment de l'intégration des outils dans une seule entité. Elle rend également possibles la structuration et l'annotation des données. En contrepartie, elle nécessite un temps de formation parfois conséquent ainsi qu'un investissement financier dans l'hypothèse de l'acquisition d'un logiciel payant.

Les plateformes analysées dans cette étude ont bien saisi l'intérêt de proposer une interface attractive et fonctionnelle permettant l'édition en ligne. C'est pourquoi elles intègrent toutes un éditeur de transcription plus ou moins complet accompagné d'une série de fonctions qui facilite l'édition et la correction des transcriptions. Ainsi, la plupart d'entre elles permettent d'enrichir les transcriptions à travers la génération et la modification de balises temporelles. Certains outils sont de plus capables de rectifier automatiquement les repères temporels lors de la césure d'un énoncé. L'identification manuelle des locuteurs et locutrices est parfois possible grâce à la disponibilité d'un champ spécifique (cf. Tableau 3). Des fonctions d'assistance à la transcription sont la plupart du temps proposées. Elles portent notamment sur la synchronisation entre le texte et le son, l'écoute et la navigation au sein de l'extrait audio (réglage de la vitesse d'écoute, « rembobinage » de quelques secondes, affichage de la forme d'onde, raccourcis clavier pour la lecture...). D'autres sont liées à l'interface (fluidité générale, simplicité d'utilisation...) ou aux caractéristiques de la transcription (mise en forme du texte, degré de certitude de la transcription...). Les plateformes qui nous ont le plus convaincus sur ces points sont dans l'ordre Sonix, Happy Scribe et Go Transcribe. Les moins efficaces dans ces opérations sont Vocalmatic et Vocapia.

Le tableau de synthèse suivant (cf. Tableau 3) recense les fonctions décrites ci-dessus et qui peuvent être recherchées pour faire le choix d'un de ces outils. Il permet de se faire rapidement une idée de leur richesse fonctionnelle.

En résumé, la présence d'un éditeur participe à l'intérêt intrinsèque de ces plateformes. Dans le cadre de notre étude, nous nous sommes particulièrement attachés (1) à vérifier la présence de fonctions permettant l'édition, l'enrichissement et la correction des transcriptions ; (2) évaluer la facilité de prise en main de l'éditeur : présentation générale, fluidité et stabilité, simplicité d'utilisation. L'examen de ces points nous a permis d'attribuer une note globale sur une échelle de 1 (la plus faible) à 5 (la plus élevée) (cf. Tableau 3)[6]. Cette note est calculée sur la base de la richesse fonctionnelle de l'outil et de sa facilité d'utilisation. Elle a été attribuée par deux membres du groupe rompus à l'utilisation d'une large gamme de logiciels dans le domaine du traitement et de l'analyse de données plurimédias. Ceci étant dit, nous insistons sur le caractère subjectif de cette note notamment en ce qui concerne l'évaluation de la présentation et la facilité d'emploi des outils, tant les ressentis peuvent être variables selon les utilisateurs et utilisatrices.

Tableau 3. Éditeurs de transcription : tableau de synthèse (Voir Tableau 9 en annexe pour plus de détails).

	Go Transcribe	Happy Scribe	Headliner	Sonix	Video Indexer	Vocalmatic	Vocapia	YouTube
Editeur de transcription en ligne (O/N)	O	O	O	O	O	O	O	O
Édition des étiquettes de locuteur (O/N)	O	O	N	O	N	N	N	N
Affichage indication proximité avec texte original (O/N)	O	O	N	O	N	N	N	N
Fonctions d'aide à la transcription (O/N)	O	O	O	O	**N**	O	O (raccourcis clavier)	O
Formats d'export de la transcription dans l'éditeur (O/N)	Word PDF TXT	Word PDF TXT SRT VTT STL HTML	VTT	Word PDF TXT SRT VTT XMLFinalcut	SRT VTT VTT XML TXT CSV	Word TXT	XML Word TXT	VTT SRT SBV
Note globale donnée à l'éditeur (1 à 5)	**4**	**4,5**	**3,5**	**4,5**	**3**	**3**	**2,5**	**3,5**

O : oui; N barré : non

Nous constatons que la qualité et les fonctionnalités des éditeurs sont très variables. Au moment où nous rédigeons ce texte, ceux de Sonix, Go Transcribe et Happy Scribe nous apparaissent comme les mieux adaptés aux principaux besoins des transcripteurs. Celui de Vocapia est actuellement en retrait du point de vue de la richesse fonctionnelle.

À ce stade, les utilisateurs et utilisatrices ayant les pratiques les plus spécifiques auront sans doute déjà orienté leur choix, éliminé certaines plateformes ou déjà écarté l'option de recourir à une plateforme de ce type, par exemple en cas de traitement de données sensibles. Pour d'autres, le point crucial est maintenant l'évaluation de la performance en termes de qualité de transcription. Dans les chapitres suivants, nous déclinons cette évaluation selon trois approches et commençons par les résultats obtenus selon une métrique « classique ».

Comparaison par calcul de distance

Mesure la plus utilisée : le WER

La métrique appelée taux d'erreur sur les mots (*Word Error Rate*, WER) est la mesure la plus couramment utilisée pour évaluer les outils de reconnaissance vocale (Bunce, 2017 ; Errattahi et al., 2018).

Le WER est dérivé de la distance de Levenshtein, en travaillant au niveau du mot plutôt qu'au niveau du caractère. La distance de Levenshtein mesure la similarité entre deux chaînes de caractères. Elle est égale au nombre minimal de caractères à supprimer, insérer, ou remplacer pour passer d'une chaîne à l'autre. Autrement dit, le WER mesure le nombre minimum de modifications de mots qui sont nécessaires pour corriger la transcription. Une correspondance parfaite donne un WER de zéro, des valeurs plus élevées indiquent une précision plus faible et donc un travail de corrections manuelles plus important (Bunce, 2017). La mesure du nombre moyen d'erreurs sur les mots prend en compte trois types d'erreurs : le pourcentage de mots remplacés (sub. ; substitution), insérés (aj. ; ajout) ou supprimés (sup. ; suppression) concernant le nombre total de mots de la référence (Nr = nombre total d'occurrences dans la référence). Le WER est donc défini comme :

$$WER = \frac{\text{sub.} + \text{sup.} + \text{aj.}}{\text{Nr}} * 100 = \frac{\text{substitutions} + \text{suppressions} + \text{ajouts}}{\text{nombre de mots de la référence}} * 100$$

Un exemple est donné ci-dessous :

1	**Transcription de référence**	La	vitamine	***	***	C	c'est	bon	pour	la	santé
2	**Transcription automatique**	La	Vie	ta	mine	C	***	bon	pour	la	santé
3	**Opérations**		***sub.***	***aj.***	***aj.***		***sup.***				

Le WER se calculera de la manière suivante :

$$WER = \frac{1 + 1 + 2}{8} = 0,5 * 100 = 50$$

Bien qu'il soit le plus utilisé, le WER présente de nombreuses lacunes (Favre et al., 2013 ; Morris et al., 2004 ; Nanjo et al., 2005 ; Park et al., 2008 ; Wang et al., 2003). Tout

d'abord, le WER est un taux qui n'a pas de limite supérieure : le nombre d'ajouts étant illimité, le pourcentage d'erreurs de mots peut être supérieur à 100 %, ce qui ne permet pas de savoir si le système est bon, mais seulement s'il est meilleur qu'un autre. En outre il ne fournit aucun détail sur la nature des erreurs, or certaines erreurs peuvent être corrigées plus facilement que d'autres (comme par exemple des erreurs de flexion : *échappé* vs *échappées*). Enfin, un taux d'erreur de mots élevé n'implique pas forcément une très mauvaise compréhension d'un texte : dans une étude portant sur un corpus d'appels de France Telecom (Wang et al., 2003), le taux mesuré d'erreur des mots atteignait 38,7 % alors que « l'erreur d'interprétation des phrases » n'était que de 12 %. Bien que des métriques alternatives aient été proposées, le WER reste néanmoins la mesure la plus efficace pour évaluer des systèmes de transcription, indépendamment de l'usage postérieur qui en est fait (Ben Jannet, 2015).

Calcul du WER et résultats

Le WER a été calculé en comparant les fichiers obtenus par les plateformes aux fichiers de transcription de référence. Il a également été calculé en comparant les fichiers de transcription manuelle fournis avec les corpus audio, à nos transcriptions de référence. Une étape de normalisation supplémentaire a ici été nécessaire, afin de ne garder du texte que les mots, sans autre signe (ponctuation, chiffres, etc.). Les traitements qui ont été opérés pour tous les fichiers (transcriptions de référence, transcriptions manuelles, transcriptions issues des plateformes) sont les suivants : suppression de la casse (tout en minuscule) ; suppression de la ponctuation et des caractères alphanumériques. La normalisation ainsi que le calcul des WER ont été réalisés avec le logiciel R. Le script R est disponible en Annexe de Tancoigne et al. (2020).

Le tableau 4 présente les résultats pour chaque fichier et chaque plateforme, sur un modèle inspiré de Bunce (2018). La première colonne présente le score WER pour la transcription manuelle. La seconde colonne présente la médiane du score obtenu pour chaque fichier. Les colonnes suivantes présentent les résultats pour chacune des plateformes. La couleur varie en fonction de la proximité du score obtenu avec la médiane (gris foncé : très supérieur à la médiane, gris clair : très inférieur à la médiane, gris : proche de la médiane). Plus le score WER est proche de 0, plus les textes comparés sont proches.

Une première observation montre que, quels que soient le type d'enregistrement et la plateforme utilisés, on observe toujours une différence notable avec la transcription manuelle : il est très clair à ce jour qu'aucun outil ne fait aussi bien qu'un transcripteur ou une transcriptrice humain·e. Néanmoins, trois grands groupes se distinguent : (1) des plateformes particulièrement performantes pour les discours préparés (texte lu ou préparé en situation de monologue) : Happy Scribe, Go Transcribe, Sonix ; (2) des plateformes particulièrement performantes pour des transcriptions de parole spontanée, avec plusieurs locuteurs (Vocapia, Video Indexer), et (3) des plateformes un peu moins bonnes que toutes les autres, tous fichiers confondus : YouTube, Headliner, Vocalmatic. Happy Scribe, Go Transcribe, Sonix, Vocapia et Video Indexer se démarqueraient ainsi en termes de comparaison purement lexicale. Les usages seraient néanmoins à différencier selon les besoins des utilisateurs et utilisatrices.

Tableau 4. Comparaison des plateformes selon 4 fichiers, par calcul du WER

	Manuelle	Médiane	Happy Scribe	Go Transcribe	Sonix	Vocapia	Video Indexer	YouTube	Headliner	Vocalmatic
Conférence / Discours	1,6	14,8	11,9	12,4	11,1	14,2	18,2	15,4	31,1	26,6
Vocabulaire spécifique	0,6	19,5	14,8	14,8	16,6	18,1	20,8	28,4	28,2	28,5
Entretien face à face	18,9	50	51,8	54,2	48,3	28,1	34	36,4	76,1	75,5
Discussion spontanée	12,7	88,4	86,2	86,2	90,5	57,6	79,8	107	222,6	244,6

Si le WER nous a permis de nous faire une idée globale de la correspondance qui existe entre nos transcriptions de référence (REF) et les transcriptions produites par les plateformes (RESULT), il ne permet pas d'estimer le temps de transcription gagné en utilisant ces outils, par rapport à une transcription manuelle intégrale : il est en effet plus simple de corriger une erreur de flexion (enlever ou ajouter un *s* de fin par exemple) que de remplacer un mot erroné, alors que ces deux erreurs seront comptabilisées de la même manière dans le WER. Nous avons donc cherché à estimer le temps qui pouvait être gagné en utilisant ces outils.

Estimation du gain de temps

Plusieurs critères peuvent jouer dans l'estimation du gain de temps, en particulier l'expérience et l'équipement de la personne réalisant la transcription, qui peuvent être très variables. Nous avons ici fait le choix de réaliser ces estimations à travers le travail d'une seule personne, en précisant son expérience et son équipement, plutôt qu'en faisant une moyenne de différentes estimations de gains de temps. Nous avons choisi cette approche afin que les lecteurs et lectrices puissent si besoin se situer par rapport à cette personne, en fonction de leur propre expérience et équipement.

Le travail a été réalisé par une chargée d'études qui travaille habituellement sur des recherches en lien avec la relation formation/emploi. Elle est habituée à retranscrire des entretiens, utilise ses dix doigts sur le clavier, et peut atteindre 43 mots/minute avec une précision de 86,81 %[7]. Elle a travaillé sans équipement adapté (pédale ou logiciel d'assistance), afin de correspondre aux conditions dans lesquelles travaillent de nombreux doctorantes et doctorants, voire chercheurs et chercheuses, et qui représentent également ses conditions de travail habituelles.

Le calcul s'est fait sur des extraits audio de 2'30" pour chaque fichier. Le temps nécessaire à la correction des fichiers bruts harmonisés a été mesuré. Ces résultats ont ensuite été comparés avec ceux obtenus en retranscrivant manuellement ces mêmes 2'30"[8]. Ces temps ont été utilisés comme base pour le calcul des gains possibles en utilisant les différentes plateformes.

Correction sans mise en forme

Le premier gain de temps calculé ne prend pas en compte le temps que prendrait la rectification de la mise en page : certains fichiers obtenus fragmentent le texte dans une présentation type « sous-titres », ce qui peut nécessiter un travail supplémentaire de nettoyage pour enlever les retours à la ligne et les sauts de lignes. Il ressort que les textes dont les erreurs de mots sont peu fréquentes, mais qui n'ont aucune ponctuation (comme dans [Conférence/Discours] et la version donnée par YouTube), nécessitent un temps de correction supérieur à ce qui a été évalué de façon qualitative (relecture humaine). Ce premier travail met en évidence une dissonance entre l'évaluation faite par une personne relectrice des fichiers obtenus en mode brut harmonisé et le temps que nécessite la remise en conformité du texte. La représentation subjective du temps à passer pour corriger les transcriptions obtenues à partir de la

Tableau 5. Durée de correction du texte sans mise en forme du texte

	Happy Scribe	Go Transcribe	Sonix	Vocapia	Video Indexer	YouTube	Headliner	Vocalmatic
[Conférence/ Discours]	0'27"	0'20"	0'17"	0'12"	2'57"	2'46"	2'23"	0'08"
[Vocabulaire spécifique]	0'16"	0'17"	0'22"	1'07"	1'41"	2'32"	2'34"	0'39"
[Entretien en face à face]	0'44"	0'00"	0'19"	0'33"	2'47"	2'04"	3'19"	0'17"
[Discussion spontanée]	0'47"	0'15"	1'31"	0'28"	2'06"	1'35"	1'18"	0'23"

Tableau 6. Durée de correction du texte avec mise en forme du texte

	Happy Scribe	Go Transcribe	Sonix	Vocapia	Video Indexer	YouTube	Headliner	Vocalmatic
[Conférence/ Discours]	8'16"	6'24"	5'25"	6'32"	10'23"	10'49"	9'17"	10'37"
[Vocabulaire spécifique]	7'36"	8'11"	7'34"	9'23"	9'54"	13'58"	10'58"	11'30"
[Entretien en face à face]	9'10"	9'9"	8'35"	6'47"	8'27"	11'15"	14'19"	11'24"
[Discussion spontanée]	13'24"	15'30"	16'33"	11'56"	14'38"	14'21"	17'26"	16'44"

lecture des fichiers, ne coïncide pas avec le temps réellement passé pour corriger les fichiers (cf. Tableau 5).

Correction avec mise en forme

Si l'on prend en compte le temps nécessaire à la remise en forme du texte (rassembler les phrases ; rajouter certaines majuscules oubliées lors de la première réécoute et corrections – Tableau 5), cela modifie sensiblement les résultats pour certaines plateformes (Tableau 6).

Ce paramètre n'est donc pas anodin et ne peut être estimé à partir du WER : l'estimation de la qualité d'une transcription ne peut se faire sans tenir compte des usages ultérieurs qui en seront faits. Il faut en outre garder à l'esprit que d'autres facteurs sont susceptibles de faire varier les résultats de gain de temps obtenus : *(1) le temps de prise en main de l'interface proposée par la plateforme* : ce temps dépend essentiellement de la simplicité d'utilisation de l'interface ainsi que de l'aisance de l'utilisateur face aux outils informatiques, *(2) le temps d'obtention des fichiers,* c'est-à-dire le temps entre le moment du dépôt du fichier multimédia sur la plateforme et la disponibilité de sa transcription.

Tableau 7. Gain de temps en % par plateforme et corpus

	Go Transcribe	Happy Scribe	Headliner	Sonix	Video Indexer	Vocalmatic	Vocapia	YouTube
[Conférence/ Discours]	**68 %**	**59 %**	**54 %**	**73 %**	48 %	47 %	**67 %**	46 %
[Vocabulaire spécifique]	42 %	46 %	22 %	46 %	29 %	18 %	33 %	0 %
[Entretien en face à face]	49 %	49 %	20 %	**52 %**	**53 %**	37 %	**62 %**	38 %
[Discussion spontanée]	30 %	39 %	21 %	25 %	33 %	24 %	46 %	35 %

En gras : valeurs supérieures à 50%.

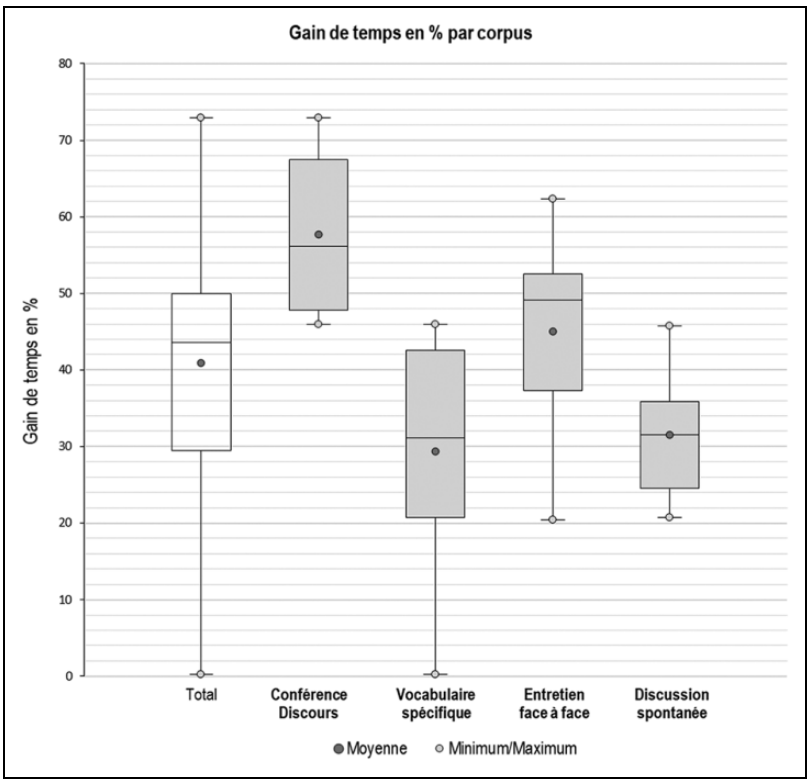

Figure 1. Gain de temps par corpus.

Le Tableau 7 présente les gains de temps en pourcentage par plateforme et par corpus. Les Figures 1 et 2 proposent une représentation sous forme de boîte à moustache de ces valeurs. Notons que ces représentations graphiques sont là pour favoriser la visualisation des résultats obtenus, mais qu'avec un échantillon de 4 sous-corpus, il semble très

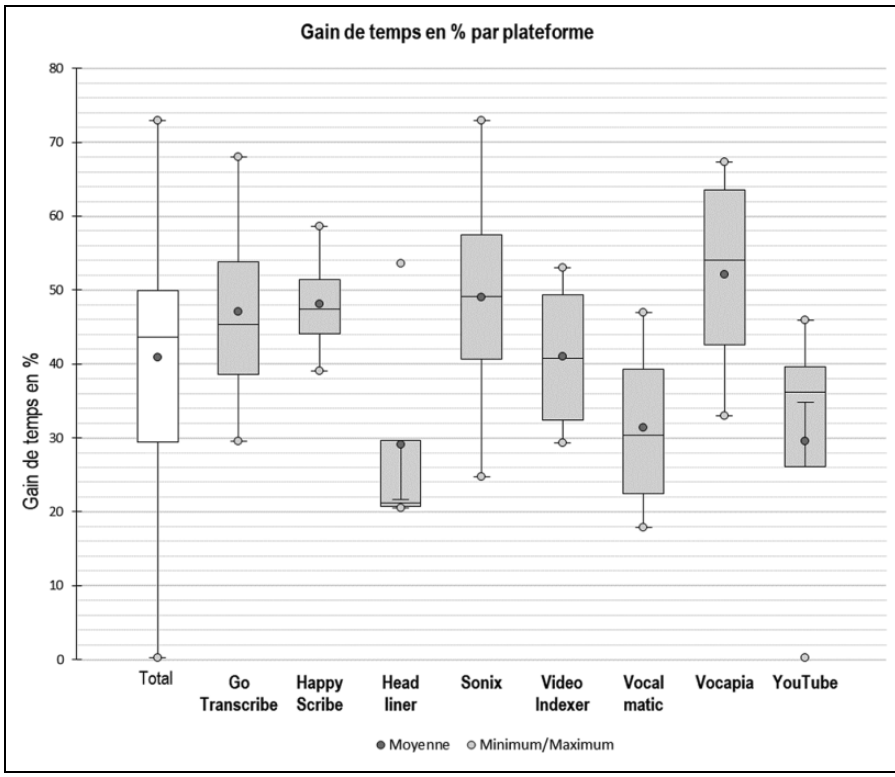

Figure 2. Gain de temps par plateforme.

prématuré d'évoquer la moindre différence significative au sens statistique du terme. Il nous semble néanmoins qu'observer des différences constatées peut aiguiller l'utilisateur ou l'utilisatrice vers tel ou tel outil selon ses contraintes propres.

Le gain de temps est variable et n'existe pas toujours (Tableau 7). Quand il existe, il est à minima de 18 % et peut monter à 73 %. Sans surprise, c'est [Conférence/Discours], le cours magistral, qui obtient le meilleur gain avec une médiane à 56 % de gain de temps. Même pour [Entretien en face à face], l'entretien sociologique avec voix atypique, le gain est notable (médiane à 49 %). [Vocabulaire spécifique] est tiré vers le bas par la (très) mauvaise prestation de YouTube sur ce fichier, qui déséquilibre la série. Enfin, sur [Discussion spontanée], la discussion légèrement cacophonique, on retrouve bien une hiérarchie entre les plateformes, mais les difficultés particulières à ce fichier se sont posées de la même façon pour l'ensemble des outils : la boite à moustache est équilibrée, dans la zone basse des gains de temps (Figure 1). On remarque tout de même que le WER de [Vocabulaire spécifique] était globalement bien meilleur que celui de [Discussion spontanée] pour – au final – aboutir aux gains de temps les plus faibles. Une étude du coût de correction des différents types d'erreurs rencontrées dans les transcriptions permettrait peut-être d'expliquer cette observation.

D'après la Figure 2, Headliner et YouTube se démarquent, avec chacune des distributions comportant des valeurs atypiques. Ils partagent également avec Vocalmatic les scores de gain de temps les plus bas, avec une moyenne avoisinant les 30 %. Deux plateformes, Happy Scribe et Video Indexer, sont plus robustes que les autres : elles semblent mieux résister à la variation liée aux différences de situation. Enfin, Vocapia offre le meilleur gain de temps tous extraits confondus.

En résumé, les classements relatifs obtenus par les deux méthodes quantitatives (calcul du WER et calcul du gain de temps) sont congruents. Cependant il n'y a pas de relation linéaire entre WER et gain de temps. Par exemple, [Vocabulaire spécifique] – Happy Scribe obtient un WER de 15 et [Entretien en face à face] – Vocapia un WER de 28 (deux fois plus élevé). Pourtant le gain de temps est de 46 % dans un cas et 62 % dans l'autre. Il n'est donc pas possible de faire des estimations de gain de temps à partir du simple calcul du WER comme cela pourrait être tentant de le faire.

WER et estimation du gain de temps : ces deux approches quantitatives sont très informatives, mais elles ne permettent pas de comprendre sur quoi butent les plateformes, et ce qui explique leurs résultats. Nous avons souhaité compléter ces analyses en réalisant un parcours systématique des transcriptions produites et des erreurs qui s'y trouvent. Ce travail repose sur la mise au point d'une typologie des erreurs introduites, comme une grille de lecture complémentaire pour choisir l'instrument adéquat pour un projet donné. Cette typologie est réalisée empiriquement, à travers la relecture exhaustive des fichiers RESULT produits pour les trois sous-corpus [Vocabulaire spécifique], [Conférence/Discours] et [Entretien en face à face][9].

Au-delà des métriques, un regard sur les transcriptions produites

Cette étape ne vise pas à établir un nouvel indicateur. Elle cherche à proposer une méthodologie pour porter un regard critique sur les objets textuels issus des transcriptions produites. Dans la lignée des réflexions menées sur l'instrumentation de la linguistique, elle invite lecteurs et lectrices à s'interroger sur les erreurs introduites et sur l'objet langagier ici manipulé. Tout comme l'annotation dont parle Habert, la transcription automatique « *n'est pas une donnée intangible, mais un résultat temporaire qu'on doit pouvoir corriger, faire évoluer* » (Habert, 2005 : 168). Une étude fine des erreurs sur un échantillon de corpus permettra de se faire une idée du coût de correction et des biais qu'une telle correction permettra d'éviter.

Contrairement à de nombreux travaux sur la question (Adda-Decker, 2006 ; Goldwater et al., 2010 ; Santiago et al., 2015 ; Errattahi et al., 2019)[10], la méthodologie proposée ici ne vise pas à identifier les erreurs produites par les plateformes en vue de participer à leur amélioration. Nous nous plaçons ici du côté de l'utilisateur ou de l'utilisatrice qui souhaite évaluer si l'état actuel de la technologie de transcription automatique implémentée dans les plateformes à sa disposition répond à ses besoins, quel type d'erreurs sont produites et quelles conséquences ces erreurs peuvent avoir sur l'analyse scientifique des données.

Afin d'amorcer cette étude, nous avons cherché un outil de mise en évidence des différences qui soit facile à prendre en main, gratuit et qui permette un accès rapide au

contenu textuel des transcriptions produites. Copyscape, un outil de détection de plagiat, permettait de répondre à ce cahier des charges. L'outil effectue une comparaison mot à mot des deux sources qui lui sont soumises. Notons que dans cette section, nous appelons **mot** tout segment textuel séparé par des espaces ou des traits d'union (*celui-ci* compte pour 2 mots, tout autant que *mais certes*). L'apostrophe, en revanche, n'est pas considérée comme un séparateur de mots (*s'enflent* compte pour un mot). Nous sommes conscients que cette définition du mot n'a aucune valeur linguistique, il s'agit d'un artefact basé sur une segmentation systématique des chaînes de caractères à partir d'une liste close de caractères délimiteurs. Copyscape parcourt les textes de manière linéaire à la recherche de passages communs et toutes les chaînes comportant au moins trois mots consécutifs identiques sont considérées comme des segments communs et placées sur une même ligne dans un tableau. Les signes de ponctuation, la casse et les sauts de ligne ne sont pas pris en compte. Cette première étape nous a permis d'accéder rapidement à deux indices de qualité lexicale des transcriptions : combien de mots comportent les transcriptions produites par chacune des plateformes et quelle proportion de celles-ci est à ce point identique à nos transcriptions de référence qu'elle est suspectée d'être du plagiat. Elle nous a également fourni un format intéressant pour servir de base à l'analyse plus fine que nous souhaitions réaliser.

Grille d'analyse des erreurs de transcription

La sortie fournie par Copyscape indique précisément les passages considérés comme ne relevant pas du plagiat. Ce sont ces passages que nous avons choisi d'annoter manuellement pour établir notre typologie. Ce travail manuel nous a permis de connaître plus en détail le contenu des transcriptions produites par les différentes plateformes.

Nous avions une idée préalable de ce que nous allions trouver, mais notre typologie s'est affinée au fur et à mesure. Elle a fait l'objet de la rédaction d'un guide d'annotation, disponible dans Tancoigne et al. (2020). Ce guide, associé à des relectures croisées, nous a permis d'uniformiser la manière dont nous utilisions chaque catégorie. Il offre un regard complémentaire à ceux des autres méthodes de comparaison des plateformes proposées dans ce document, une analyse qualitative aux côtés de deux approches quantitatives. Nous invitons les lecteurs et lectrices à consulter Tancoigne et al. (2020) pour s'approprier ce regard et adapter notre typologie en fonction de leurs besoins spécifiques. La typologie d'analyse des mots des transcriptions, illustrée par des erreurs réellement observées, est la suivante :

- Mots communs → *quasi* transcrit *quasi*
- Erreurs de flexion → *épineuses* transcrit *épineuse*
- Substitution avec proximité phonétique → *l'homme* transcrit *Lomme*
- Substitution sans proximité phonétique → *mairie* transcrit *vallée*
- Passage sans lien → *est pile à cheval* transcrit *je vais le faire*
- Mots de REF absents de RESULT (suppression) → *saine* non transcrit
- Mots de RESULT qui ne correspondent à rien dans REF (ajout) → *sont inductions* transcrit *sont **à** induction*
- Autres → *point* transcrit à l'aide du signe de ponctuation.

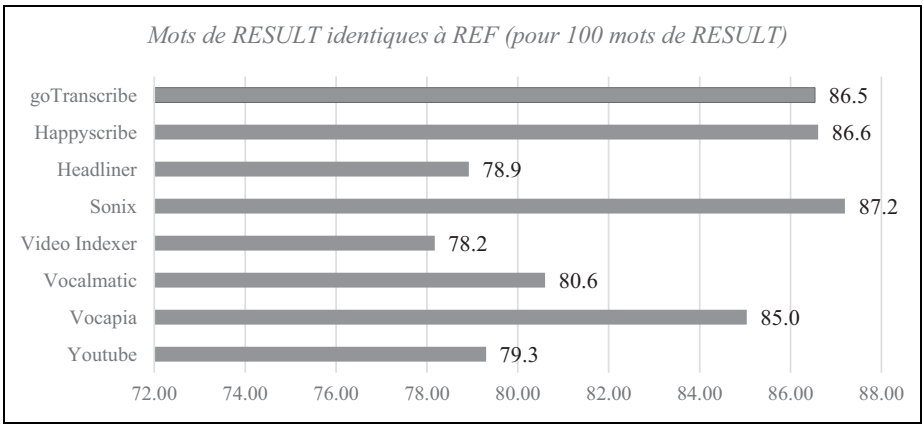

Figure 3. Proportion de mots identiques à REF dans chaque fichier RESULT – [Entretien en face à face].

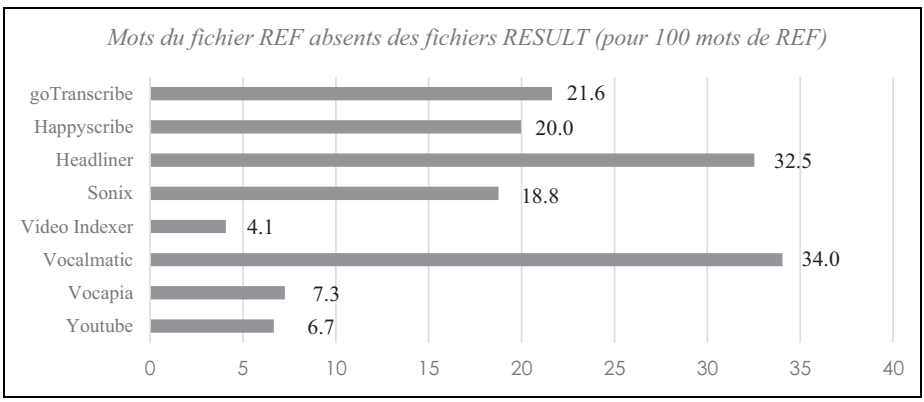

Figure 4. Proportion de mots de REF considérés comme n'ayant pas été transcrits – [Entretien en face à face].

Analyse des résultats

Trois sous-corpus ont été annotés selon cette méthode, soit un total de 24 fichiers RESULT[11]. Nous présenterons ici, pour illustration, quelques résultats issus de l'analyse du sous-corpus [Entretien en face à face]. Nous rappelons qu'il s'agit d'un extrait d'un entretien de recherche entre deux locuteurs, dont l'un présente une voix atypique (légèrement rauque et aux tonalités variables). Les conditions d'enregistrement sont relativement bonnes, sans bruit parasite.

L'ensemble des plateformes testées retourne des performances raisonnables pour le sous-corpus [Entretien en face à face], comme le montre la proportion de mots des fichiers RESULT qui se trouvent à l'identique dans le fichier REF (Figure 3), variant de 78,2 % à 87,2 %.

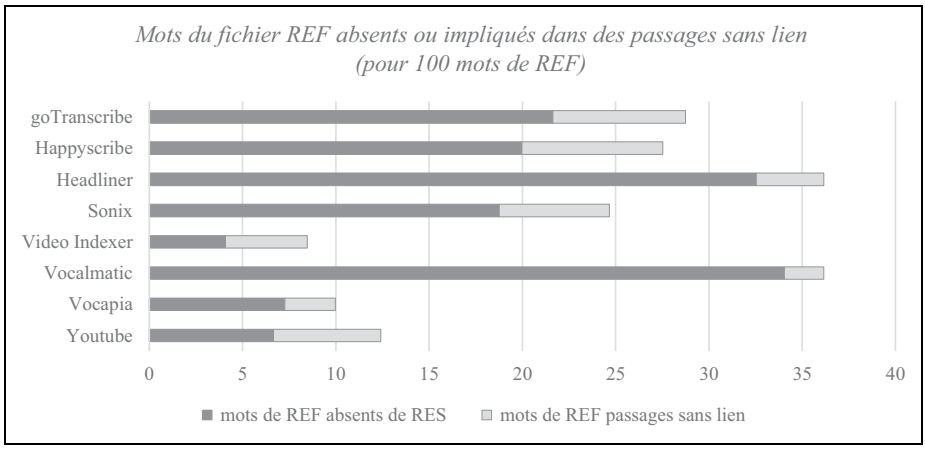

Figure 5. Proportion de mots de REF considérés non transcrits ou impliqués dans des passages sans lien – [Entretien en face à face].

Cependant, ces chiffres attractifs doivent être relativisés en regardant la taille des transcriptions produites et, par conséquent, la proportion des mots de REF pour lesquels nous avons considéré que les différents instruments n'avaient rien proposé du tout. La Figure 4 montre en effet que la couverture est très variable selon les plateformes. On distingue ici trois groupes : Vocalmatic et Headliner, qui ont produit des textes dont la taille en nombre de mots est plus de 30 % inférieure à celle du texte de référence ; Sonix, Happy Scribe et Go Transcribe, pour qui la perte se situe aux alentours de 20 % ; YouTube, Vocapia et Video Indexer enfin, qui proposent une transcription pour au moins 92,7 % des mots.

Ces trois groupes correspondent-ils à trois stratégies de transcription différentes en cas de passages présentant des difficultés particulières ? Notre grille d'analyse distingue deux types de mots de REF absents : les cas où les instruments n'ont produit aucune transcription (mots de REF absents de RESULT) et ceux où les instruments ont produit des segments textuels sans aucun lien avec la transcription de référence[12] (mots de REF passages sans lien). Cette distinction fournit un indice sur les stratégies employées. La Figure 5 montre que la proportion de mots de chacune de ces catégories varie selon les plateformes.

On retrouve dans cette figure les deux premiers groupes que nous avions identifiés :

- Vocalmatic et Headliner, qui ont produit des textes dont la taille en nombre de mots est inférieure à 70 % de celle du texte de référence, produisent peu de passages dits « sans lien » (respectivement 3 et 4 passages pour 2,12 % et 3,63 % des mots transcrits). On peut donc considérer que, pour ce sous-corpus en tout cas, la stratégie de ces instruments est de privilégier une transcription peu couvrante, en espérant qu'elle soit la plus fiable possible.
- Sonix, Happy Scribe et Go Transcribe, pour qui la perte se situe aux alentours de 20 % des mots de REF, produisent beaucoup de passages « sans liens » (respectivement 8, 8 et 10 pour 5,9 %, 7,56 % et 7,11 % des mots transcrits). On peut donc difficilement se faire une idée de la stratégie de ces plateformes à partir de nos observations.

Les trois dernières plateformes : YouTube, Vocapia et Video Indexer, qui proposent une transcription pour au moins 96 % des mots, ne semblent pas partager un comportement uniforme. Tandis que les transcriptions proposées par YouTube et Video Indexer comportent de nombreux passages « sans liens » (respectivement 10 et 8, pour 5,75 % et 4,39 % des mots transcrits), la transcription proposée par Vocapia en comporte 6, pour seulement 2,72 % des mots transcrits. On peut alors considérer que, pour ce sous-corpus en tout cas, la stratégie de YouTube et Video Indexer est de transcrire le plus possible, quitte à produire des énoncés sans rapport avec le signal. En revanche, Vocapia parvient à proposer ici une transcription couvrante de bonne qualité.

En conclusion, les transcriptions du sous-corpus [Entretien en face à face] proposées par les différentes plateformes sont de taille et de qualité variable. Elles semblent toutes nécessiter une relecture attentive, accompagnée d'un retour à la version audio du sous-corpus. La plateforme Vocapia semble tirer son épingle du jeu, en proposant une transcription à la fois couvrante et d'assez bonne qualité, conforme au gain de temps estimé de 62 % dans le Tableau 7.

La même analyse, conduite sur le sous-corpus [Vocabulaire spécifique], a mis en avant un nombre important d'erreurs de flexion. Nous avons notamment pu noter que l'antéposition de certains adjectifs épithètes, telle que *de plus vulgaires et connues actions des hommes*, pose problème à l'ensemble des plateformes qui ne parviennent pas à accorder *connues* à *actions*. De plus, les temps verbaux propres à l'écrit (*n'eussions*, *saurions*, *prissassent*) et le vocabulaire vieilli (*cettui*, *onques*) ne semblent pas connus des plateformes et provoquent des erreurs. Cela montre que, sur certaines données, l'utilisateur ou l'utilisatrice pourra optimiser les résultats à moindre coût, en intégrant la transcription à une chaîne de traitement automatique plus large. Ils pourront ainsi collecter des lexiques en amont et appliquer une correction orthographique en aval.

À l'inverse, l'étude systématique des erreurs produites pour la transcription du sous-corpus [Conférence/Discours] nous a montré que la marge de manœuvre pour améliorer automatiquement les performances des plateformes est parfois limitée et que seule une relecture manuelle fine peut permettre d'identifier et de corriger certaines erreurs (*à travers la notion d'espace* transcrit *à travers la notion d'espace**s***).

Les différentes catégories que nous avons proposées permettent de s'interroger sur les possibilités d'intégrer des lexiques supplémentaires en entrée de la transcription automatique ou d'imaginer des post-traitements pour améliorer les performances des différentes plateformes. Ces catégories permettent aussi de différencier les résultats obtenus dans certains cas, comme le sous-corpus [Entretien en face à face]. Elles ne sont pas nécessairement suffisantes pour répondre à tous les cas de figure et la typologie proposée devra être complétée selon les besoins et la nature des observations menées. Tout comme la mesure du gain de temps, elles sont à mettre au regard de ses propres objectifs de recherche et de sa discipline scientifique. L'absence de transcription des marques d'hésitations n'aura ainsi pas les mêmes conséquences selon que l'on s'intéresse à l'analyse des pratiques langagières ou à l'enchaînement logique des événements d'un récit.

Conclusion

Quelle plateforme utiliser ?

Ce travail a permis de montrer qu'il n'existe pas de plateforme qui serait performante sur tous les critères, mais plutôt des groupes de plateformes spécialisées dans des types de discours particuliers. Sonix et Vocapia[13] ressortent particulièrement bien, tant en matière de richesse fonctionnelle que de qualité des résultats obtenus : le premier plutôt pour les discours planifiés, le second pour le traitement de la parole spontanée. Malgré les nombreuses recherches menées au cours des 10 à 15 dernières années pour améliorer la précision des systèmes automatiques en corrigeant des erreurs de transcription – et même si les résultats sont prometteurs – pour l'heure, la majorité des outils nécessitent au final une correction manuelle des fichiers obtenus. Ces outils pourraient néanmoins faire gagner jusqu'à 75 % du temps de travail nécessaire à une transcription intégrale des enregistrements audio, pour un usage de type entretien sociologique.

Vers une offre académique ?

Nous l'avons vu, la plupart de ces outils posent des problèmes en termes de confidentialité des données. Ceci plaide pour la construction d'une offre académique, libre et hébergée sur le territoire national. Le potentiel d'utilisation d'un tel outil par des établissements universitaires est immense : non seulement en termes d'utilisation dans le cadre de recherches (il serait par ailleurs utile de quantifier les usages de la transcription à cette fin), mais également en termes d'enseignement. Le déploiement d'une offre de formation en ligne de plus en plus importante dans les années à venir nécessitera le recours à une utilisation massive d'outils d'indexation de contenus ainsi que de sous-titrages. Il permettrait également le déploiement de ces cours à un public à l'étranger, notamment avec un couplage à une offre de traduction. Le développement d'une telle offre pourrait changer le visage du marché en constante évolution de la transcription automatique, actuellement majoritairement privé.

Une méthode pertinente ?

La méthode d'évaluation présentée dans ce travail possède ses limites, notamment en ce qui concerne le choix des corpus ou encore celui de faire appel à l'expertise d'une seule personne pour l'évaluation du gain de temps. Sa mise en œuvre requiert en outre des ressources importantes en temps et en compétences, alors même que les résultats qu'elle permet d'obtenir ne sont valables que pour un temps donné. Malgré ces limites, elle reste beaucoup plus complète que tout ce qui a pu être fait jusqu'à présent et nous pensons qu'elle possède un certain nombre d'avantages pour qui souhaite réaliser une évaluation de ce type d'outils dans le cadre de son travail de recherche. D'une part, elle permet de sélectionner, parmi un certain nombre de critères déjà définis, ceux qui seront particulièrement importants pour une recherche donnée, et d'orienter ainsi les efforts d'évaluation sur ces critères. D'autre part, elle permet de relativiser l'importance qui peut être donnée au WER, et suggère de concentrer les efforts d'évaluation directement sur l'évaluation du temps qui peut être gagné pour un extrait audio particulier. Enfin, elle

encourage avant tout et surtout à réaliser des essais *in vivo*, c'est-à-dire de tester soi-même de cours extraits de ses propres audio de recherche sur différentes plateformes. Ce travail d'évaluation et de comparaison, nécessaire si l'on souhaite faire appel à ce type d'outil, pourrait être planifié dans l'économie des projets de recherche afin de pouvoir bénéficier de ressources dédiées (temps, crédits de transcription). Il serait intéressant que de tels essais contribuent par la suite à enrichir notre étude, à la fois en termes de corpus, de tests et de regards disciplinaires différents.

Déclaration de conflits d'intérêts

Les auteur.es déclarent n'avoir aucun conflit d'intérêt potentiel pour tout ce qui concerne le déroulement de la recherche, les droits d'auteur et/ou la publication de cet article.

Financement

Les auteur.es n'ont bénéficié d'aucun soutien financier particulier relatif au déroulement de la recherche, à leurs droits d'auteur et/ou à la publication de cet article.

Remerciements

Nous tenons tout d'abord à remercier le Comité MATE-SHS, le Comité d'organisation des Tuto@MATE pour son invitation à présenter une première version de ce travail, et tou.te.s les participant.e.s de cette session pour leurs intérêt, questions et commentaires. Nous remercions également les membres des projets MONLOE et TCOF, grâce à qui nous avons pu travailler sur des corpus et des transcriptions de qualité, ainsi que Camille, pour s'être prêté au jeu de l'entretien dans le cadre de ce projet. Merci également à nos différentes tutelles, qui nous ont permis de consacrer du temps à ce projet hors cadre et nous ont fourni les ressources logicielles nécessaires pour le mener à bien. Enfin, merci aux relecteurs et à l'équipe du *BMS* dont les suggestions nous ont permis d'améliorer la structuration de cet article.

Notes

1. https://mate-shs.cnrs.fr/ Le réseau compte environ 500 membres, partout en France.
2. Cette durée peut a priori paraître courte. Cependant, elle se situe dans la moyenne des études préexistantes, qui privilégient généralement des échantillons de 4 à 8 minutes. Elle permet en outre de pouvoir effectuer les tests sur les plateformes payantes sans souscrire d'abonnement (les temps d'essais sont généralement limités à 30 minutes, ce qui permet de tester quatre extraits de cinq minutes).
3. Corpus « comptines_cou_10 » disponible sur https://tcof.atilf.fr/
4. Disponible sur https://montaigne.univ-tours.fr/category/multimedia/ed-sonore/
5. Coût calculé sur la base du SMIC horaire brut, au 30 mai 2020 et en prenant en compte des charges patronales à hauteur de 34 %.
6. Des copies d'écran commentées sont disponibles dans Tancoigne et al. (2020).
7. Test réalisé sur http://typing-test.arckade.fr/
8. Pour chaque fichier le temps de transcription des 2'30" est le suivant : [Vocabulaire spécifique] : 14' ou 840" ; [Conférence/discours] : 20' ou 1200", [Entretien en face-à-face] : 18' ou 1080", [Discussion spontanée] : 22' ou 1320".
9. Le sous-corpus [Discussion spontanée] a été exclu au cours de la procédure en raison du temps de traitement induit par sa complexité : nombreux chevauchements de parole, passages

inaudibles se traduisant par des transcriptions lacunaires, parcellaires et/ou fantaisistes. Son analyse aurait nécessité la mise au point d'une méthodologie ad hoc.

10. Nous vous invitons à consulter ces travaux pour en apprendre davantage sur le fonctionnement de la transcription automatique et les phénomènes langagiers en jeu dans les erreurs produites.
11. Chaque corpus correspond à un sous-ensemble de fichiers, composés d'une transcription de référence (fichier REF) et huit transcriptions automatiques (fichiers RESULT).
12. Exemple de passage sans lien : *est pile à cheval* dans REF transcrit *je vais le faire* par une plateforme.
13. Une instance du moteur de reconnaissance de Vocapia est disponible sur la plateforme « Yobi Yoba ». Cette plateforme, apparue après notre étude, dispose d'une interface riche et fonctionnelle.
14. Portion de code informatique permettant d'utiliser les plateformes sans passer par l'interface Web
15. Transcription des marques de formulation et disfluences : euh, hein, bah, ben, hum..., répétitions, bégaiements, etc.

Références

Authôt (2016) Reconnaissance automatique de la parole au coeur de l'application Authôt. Voir : https://www.authot.com/fr/2016/09/09/systeme-reconnaissance-de-la-parole/ (accédé le 3 juillet 2020)

Adda-Decker M (2006) De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. *In* : *JEP-TALN 2006– Journées d'Études sur la Parole et conférence annuelle du Traitement Automatique des Langues Naturelles*. Dinard, 877-888.

Baude O, Blanche-Benveniste C, Calas M.-F, Cappeau P, Cordereix P, Goury L, Jacobson M, De Lamberterie I, Marchello-Nizia C et Mondada L (2006) *Corpus oraux. Guide des bonnes pratiques*. CNRS Editions.

Ben Jannet M.A (2015) Évaluation adaptative des systèmes de transcription en contexte applicatif. *These de doctorat, Université Paris-Saclay (COMUE)*.

Blanche-Benveniste C (2000) Corpus de français parlé. In : Bilger M. (éd.) *Corpus : Méthodologie et applications linguistiques*. Paris : Honoré Champion, 15-25.

Benzitoun C, Fort K et Sagot B (2012) TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. *In* : *JEP-TALN 2012– Journées d'Études sur la Parole et conférence annuelle du Traitement Automatique des Langues Naturelles*. Grenoble, 99-112.

Bunce T (2017) Comparing Transcriptions, *Not this*...Voir : https://blog.timbunce.org/2017/02/09/comparing-transcriptions/ (accédé le 4 juillet 2020).

Bunce T (2018) A Comparison of Automatic Speech Recognition (ASR) Systems, *Not this*...Voir : https://blog.timbunce.org/2018/05/15/a-comparison-of-automatic-speech-recognition-asr-systems/ (accédé le 3 juillet 2020).

Bunce T (2020) A Comparison of Automatic Speech Recognition (ASR) Systems, part 3, *Not this*... Voir : https://blog.timbunce.org/2020/05/17/a-comparison-of-automatic-speech-recognition-asr-systems-part-3/ (accédé le 3 juillet 2020).

Casilli AA (2019) *En attendant les robots : enquête sur le travail du clic*. Paris XIXe : Éditions du Seuil.

Cintas JD et Remael A (2014) *Audiovisual translation : subtitling*. London : Routledge.

Dacos M et Mounier P (2015) Humanités numériques. État des lieux et positionnement de la recherche française dans le contexte international. Paris : Institut Français.

Errattahi R, El Hannani A et Hain T (2019) System-independent ASR error detection and classification using Recurrent Neural Network, *Computer Speech & Language*, 55: 187-199.

Errattahi R, El Hannani A et Ouahmanea H (2018) Automatic Speech Recognition Errors Detection and Correction: A Review, *Procedia Computer Science*, 128: 32-37.

Favre B, Cheung K, Kazemian S, Lee A, Liu Y, Munteanu C, Nenkova A, Ochei D, Penn G, Tratz S, Voss C et Zeller F (2013) Automatic human utility evaluation of ASR systems: Does WER really predict performance ? *In* : *Proceedings of Interspeech 2013*. Lyon, 3463-3467.

Goldwater S, Jurafsky D et Manning C.D (2010) Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates, *Speech Communication*, 52(3): 181-200.

Habert B (2005) Portrait de linguiste (s) à l'instrument. In Guillot C, et al. (éd.), *À la quête du sens : études littéraires, historiques et linguistiques en hommage à Christiane Marchello-Nizia*. Lyon : ENS Editions, 163-174.

Heiden S, Magué J.-P et Pincemin B (2010). TXM : Une plateforme logicielle open-source pour la textométrie– conception et développement. In : *JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data* Rome, Italie.

Hitlin P (2016) Turkers in this canvassing: young, well-educated and frequent users, *Research in the Crowdsourcing Age, a Case Study*. Voir : https://www.pewresearch.org/internet/2016/07/11/turkers-in-this-canvassing-young-well-educated-and-frequent-users/ (accédé le 22 septembre 2021).

Këpuska V (2017) Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx), *International Journal of Engineering Research and Applications*, 7(3): 20-24.

Khamsi R (2019) Say What? A Non-Scientific Comparison of Automated Transcription Services, *The Open Notebook*. Voir : https://www.theopennotebook.com/2019/12/17/say-what-a-non-scientific-comparison-of-automated-transcription-services/ (accédé le 4 octobre 2020).

Kim J.Y, Liu C, Calvo R.A, McCabe K, Taylor S.C.R, Schuller B.W et Wu K (2019) A Comparison of Online Automatic Speech Recognition Systems and the Nonverbal Responses to Unintelligible Speech, *arXiv:1904.12403 [cs.SD]*.

Kong X, Choi J.Y et Shattuck-Hufnagel S (2017) Evaluating automatic speech recognition systems in comparison with human perception results using distinctive feature measures. *In : 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA, USA, 5810-5814.

Mondada L (2000) Les effets théoriques des pratiques de transcription, *Linx. Revue des linguistes de l'Universite Paris X Nanterre*, 42: 131-146.

Mondada L (2008) La transcription dans la perspective de la linguistique interactionnelle. In : Bilger M. (éd.) *Données orales. Les enjeux de la transcripion. Perpignan* : Presses Universitaires de Perpignan, 78-110.

Morris A, Maier V et Green P (2004) From WER and RIL to MER and WIL : improved evaluation measures for connected speech recognition. *In : Proceedings Of Interspeech 2004*. Jeju Island, Korea, 2765-2768.

Nanjo H et Kawahara T (2005) A New ASR Evaluation Measure and Minimum Bayes-Risk Decoding for Open-domain Speech Understanding. In : *Proceedings of ICASSP '05, IEEE International Conference on Acoustics, Speech, and Signal Processing*. Philadelphia, 1053-1056.

Park Y, Patwardhan S, Visweswariah K et Gates S.C (2008) An empirical analysis of word error rate and keyword error rate. In : *Proceedings of Interspeech 2008*. Brisbane, 2070-2073.

Rioufreyt T (2016) La transcription d'entretiens en sciences sociales : Enjeux, conseils et manières de faire. Halshs-01339474

Rioufreyt T (2019) Usage et appropriation des logiciels d'aide à l'analyse qualitative. *Bulletin de méthodologie sociologique*, 143(1): 7-27.

Santiago F, Dutrey C et Adda-Decker M (2015) Towards a typology of ASR errors via syntax-prosody mapping. *In* : Adda G., et al. (éd.) : *Proceedings of ERRARE 2015*. Sinaia, Romania : Editura Academiei Române, 175-192.

Schmid H (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. In : *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK.

Tancoigne E, Corbellini J.-P, Deletraz G, Gayraud L, Ollinger S et Valero D (2020) La transcription automatique : un rêve enfin accessible ? Analyse et comparaison d'outils pour les SHS. Nouvelle méthodologie et résultats. *Rapport de recherche, MATE-SHS*.

Tardieu J (1951) *Un Mot Pour Un Autre*. Paris : Gallimard.

Vuylsteker B (2017) Speech Recognition — A comparison of popular services in EN and NL, *Craftworkz*. Voir : https://blog.craftworkz.co/speech-recognition-a-comparison-of-popular-services-in-en-and-nl-67a3e1b0cee6 (accédé le 4 octobre 2020).

Wang Y-Y, Acero A et Chelba C (2003) Is word error rate a good indicator for spoken language understanding accuracy. In : *2003 IEEE Workshop on Automatic Speech Recognition and Understanding* (IEEE Cat. No.03EX721). St Thomas : 577-582.

Annexe: tableaux récapitulatifs des fonctionnalités des plateformes

Tableau 8. Caractéristiques générales des plateformes (mars 2020)

	Go Transcribe	Happy Scribe	Headliner	Sonix	Video Indexer	Vocalmatic	Vocapia	YouTube
Présence CGU ou mentions légales (O/N)	O	O	O	O	O	O	O	O
Langue des documents	Anglais	Anglais	Anglais	Anglais	Anglais	Anglais	Anglais	Français
RÈGLEMENTATIONS APPLIQUÉES								
Nom de la société	Go-Transcribe Ltd	Happy Scribe Ltd	SpareMin	Sonix Inc.	Microsoft	Enactics Inc.	Recher-che Vocapia - Vocapia Research	(pour l'espace européen) Google Ireland Ltd (groupe de sociétés Alphabet Inc.)
Siège social	Londres Angleterre	Dublin Irlande	New York États-Unis	San Francisco États-Unis	Redmond États-Unis	Ontario Canada	Orsay France	Dublin Irlande
Soumis à la RGPD (O/N)	O	O	O (extra-territoria-lité)	O (extra-territoria-lité)	O (extra-territoria-lité)	O (extra-territoria-lité)	O	O (extra-territoria-lité)
Hébergeur des données	Microsoft	Amazon Google	N/D	N/D	Microsoft	Google	N/D	Google
Réglementation nationale appliquée (RGPD, Cloud Act, Patriot Act . . .)	RGPD	RGPD	États-Unis, Californie	États-Unis	« Microsoft Data Subjects Rights » et CCPA– « California Consumer Privacy Act »	Canada	RGPD	RGPD et EU-U.S. et Swiss-U.S. Privacy Shield Frame-works

(à suivre)

Tableau 8. (suite)

	Go Transcribe	Happy Scribe	Headliner	Sonix	Video Indexer	Vocalmatic	Vocapia	YouTube
PROTOCOLES DE CRYPTAGE								
Protocole d'échange des fichiers (HTTPS . . .)	HTTPS	HTTPS	HTTPS	HTTPS	HTTPS	HTTPS	HTTPS	HTTPS
Chiffrement du transfert des données (protocole et longueur de clé)	TLS 256 bits	TLS 128/256 bits	TLS 128/256 bits	TLS 128/256 bits	TLS 256 bits	TLS 256 bits	TLS 128/ 256 bits	TLS 128/ 256 bits
Autorité de certification	Let's Encrypt Authority X3	Let's Encrypt Authority X3	Amazon	Cloudflare	Microsoft Corporation	Let's Encrypt Authority X3	Sertigo GB	Google Trust services
Tarification des services de transcription								
Tarifs services payants pour une heure	8 à 12 €	9,60 et 12 € par heure	Service « pro » 12,95 $ par mois ou 120 $ par an ; service « pro »gratuit pour usage académi-que	5 et 10 €	9 € pour la vidéo 2,40 € pour l'audio Tarification à la minute. En fonction des services demandés.	6 à 15 €	5 et 10 €	Plateforme gratuite
Tarifs académiques (O/N)	N	N/D	O (gratuité totale du service « pro »)	O (jusqu'à 50 % de remise sur tarif public)	N/D	O dans le cadre d'une offre personna-lisée	Sur demande	N/A
Services gratuits, période d'essai non compris (O/N)	N (30 mn d'essai)	N (30 mn d'essai)	O 10 fichiers par mois (au-dessus présence de fili-grane), 350 Mo max. par fichier	N (30 mn d'essai)	O (10 heures de média ; 40 heures à travers l'API[14])	N (30 mn d'essai)	N (essai sur demande)	O Plateforme gratuite

(à suivre)

Tableau 8. (suite)

	Go Transcribe	Happy Scribe	Headliner	Sonix	Video Indexer	Vocalmatic	Vocapia	YouTube
Système de parrainage (O/N)	O	O	O	O	N	O	N	N/A
Caractéristiques et métadonnées des transcriptions								
Langues supportées	Principales langues occidenta-les et slaves. Japonais	Principales langues occidenta-les, arabes, asiatiques et slaves	Principales langues occidenta-le, arabes, asiatiques et slaves	Principales langues occidenta-les, arabes, asiatiques et slaves	Principales langues occidenta-les, arabes, asiatiques et slaves	Principales langues occidenta-les, arabes, asiatiques et slaves	Principa-les langues occidenta-les, arabes, asiatiques et slaves, hébreu, hindi, pachto, swahili, ourdou	Allemand, anglais, russe, coréen, italien, espagnol, français, japonais, néerlan-dais, portugais
Gestion des codes temporels (O/N)	O	O	O	O	O	O	O	O
Détection automatique changement de locuteur (O/N)	O	O	N	O(en version béta)	N/D	N	O	N
Identification automatique des locuteurs (O/N)	N	N	N	N	N	N	O	N
Étiquetage locuteurs (O/N)	O	O	N	O	N	N	O	N
Transcription de la ponctuation (O/N)	O	O	O	O	O	N	O	N
Enrichissement du lexique (O/N)	N	O(vocabu-laire spécifique) 100 termes max.	N	O(vocabu-laire spécifique)	O(personna-lisation des modèles)	N	O(personnalisation des modèles sur demande)	N

(à suivre)

Tableau 8. (suite)

	Go Transcribe	Happy Scribe	Headliner	Sonix	Video Indexer	Vocalmatic	Vocapia	YouTube
Marques d'élocution[15] (O/N)	N	N	N	N	Partielle	N	Partielle	Partielle
Formats de fichiers en entrée et en sortie								
Taille et durée maxi	1 Go	N/D	350 Mo ou 1 heure ; audio max 2 heures— 500 Mo	4 Go	2 Go	N/D	4 Go ou 4 heures	15 min ou 12 heures/128 Go après validation du compte
Formats audio en entrée	WAV, MP3, m4a, AAC	AAC, AIFF, WAV, MP3, ASF, FLAC . . .	WAV, MP3	AAC, AIFF, WAV, MP3, ASF, FLAC . . .	MP2, FLAC, WAV	AAC, AIFF, ASF, FLAC, WAV, MPEG, Ogg Vorbis . . .	AAC, MP3, WMA, WAV, FLAC, Opus, Vorbis, AMR	Aucun
Formats vidéo en entrée	AVI, MP4, MOV	AVI, MP4, MOV, FLV, Mpeg, WMV . . .	MP4, MOV	AVI, MP4, FLV, MOV, WMV . . .	MP4, MOV, OGG, Webm (vidéo acceptée, mais transcrip-tion limitée à l'anglais)	Aucun	AVI, MOV, MP4, MPEG2, 3GP, WMV, WebM, MKV . . .)	AVI, MP4, MOV, WMV, MPG, FLV, 3GPP, WebM, DNxH, ProRes CineForm, HEVC (h265)
Formats transcription texte en sortie	DOC, TXT, PDF, SRT, VTT	DOC, PDF, HTML, TXT, VTT, SRT, STL, Première	VTT, MP3, MP4 (incrusta-tion transcrip-tion)	DOC, TXT, SRT, VTT FCPxml, xmlPremi-ère	SRT, VTT, TTML, TXT, CSV	DOC, TXT	XML, VTT, RTF, PDF	SRT, VTT, SBV

O : oui ; N : non ; N/A = non adapté ; N/D = non determiné

Tableau 9. Éditeur de transcription : description

	Go Transcribe	Happy Scribe	Head liner	Sonix	Video Indexer	Vocalmatic	Vocapia	You Tube
Editeur de transcription en ligne (O/N)	O	O	O	O	O	O	O	O
Édition des étiquettes de locuteur	O	O	N	O	N	N	N	N
Formats d'exportation de la transcription dans l'éditeur	DOC, PDF, TXT	DOC, PDF, TXT, SRT, VTT, STL, HTML	VTT	DOC, PDF, TXT, SRT, VTT, XML Finalcut	SRT, VTT, TTML, TXT, CSV	DOC, TXT	DOC, XML, TXT	SBV
Balises temporelles								
Modification manuelle de la segmentation des balises temporelles (O/N)	O	O	O	O	N	N	N	O
Réalignement automatique des balises temporelles après segmentation de l'énoncé (unité de segmentation) (O/N)	O(mot)	O(mot)	N/A	O(mot)	N/A	N/A	N/A	O(mot)
Réajustement du temps de départ (O/N)	O	O	N	O	N/A	N/A	N	O
aide à la transcription								
Affichage indication proximité avec texte original (O/N)	O	O	N	O	N	N	N	N
Fonctions d'aide à la transcription (O/N)	O	O	O	O	N	O	O	O

O : oui ; N : non ; N/A = non adapté ; N/D = non determiné

Tableau 10. Fonctionnalités supplémentaires

	Go Transcribe	Happy Scribe	Head liner	Sonix	Video Indexer	Vocal matic	Vocapia	YouTube
[14]API disponible (O/N)	N	O	O	O	O	N	O	N
Traitement par lot (O/N)	N	O	O	O (selon abonnement)	N	N/D	O (service en ligne)	N
Fonctions collaboratives (O/N)	N	O	N	O (selon abonnement)	O	N	O	O

O : oui ; N : non ; N/D = non déterminé