La transcription automatique : un rêve enfin accessible ? Analyse et comparaison d'outils pour les SHS. Nouvelle méthodologie et résultats

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

MATE-SHS

La transcription automatique : un rêve enfin accessible ?

Analyse et comparaison d'outils pour les SHS. Nouvelle méthodologie et résultats

Élise TANCOIGNE[1], Jean-Philippe CORBELLINI[2], Gaëlle DELETRAZ[3], Laure GAYRAUD[4], Sandrine OLLINGER[5], Daniel VALERO[6]

Avec les contributions de : **Laurie Boyer[6], Justine Lascar[6], Marick Fèvre[7], Giancarlo Luxardo[8]**

Septembre 2020

atilf ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE

MSH VAL DE LOIRE

Centre Associé Régional de Bordeaux Céreq
Centre Emile Durkheim
Sciences Po Bordeaux
CNRS UMR 5116 - Université de Bordeaux

icar
UMR 5191 - CNRS / Université Lyon 2 / ENS de Lyon
Interactions, Corpus, Apprentissages, Représentations

UNIVERSITÉ DE GENÈVE

UMR 5319 Passages

[1] Université de Genève, elise.tancoigne@unige.ch
[2] Maison des Sciences de l'Homme Val-de-Loire USR 3501, jean-philippe.corbellini@univ-tours.fr
[3] UMR 5319 Passages - CNRS/Université de Pau et des Pays de l'Adour, gaelle.deletraz@univ-pau.fr
[4] Centre régional associant le Céreq au Centre Emile Durkheim - IEP de Bordeaux, l.gayraud@sciencespobordeaux.fr
[5] ATILF - UMR 7118 CNRS/Université de Lorraine, Sandrine.Ollinger@atilf.fr
[6] I.C.A.R - UMR 5191 CNRS/ENS de Lyon/Université Lyon 2, Daniel.Valero@ens-lyon.fr
[7] CENS - UMR 6025 CNRS/Université de Nantes - Réseau doctoral EHESP
[8] UMR 5267 Praxiling - CNRS/Université Paul Valéry Montpellier 3

Résumé

Le recueil de la parole est au cœur des démarches de recherches qualitatives 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 de l'intégralité de la parole enregistrée est devenue quasiment la norme, mais elle demande beaucoup de temps et s'avère souvent fastidieuse et un peu décourageante. À l'heure de l'intégration de modules d'intelligence artificielle aux algorithmes de reconnaissance automatique de la parole, ces derniers progressent rapidement et le fantasme de pouvoir automatiser cette tâche longue et pénible semble se rapprocher, voire être déjà accessible.

Ce rapport présente le résultat d'un travail de comparaison de 8 outils de transcription automatique (Go Transcribe, Happy Scribe, Headliner, Sonix, Video Indexer, Vocalmatic, Vocapia, YouTube) effectué par des membres du réseau méthodologique CNRS MATE-SHS. Quatre extraits de fichiers audio de langue française ont servi de test, chacun avec ses spécificités propres : un texte lu, un cours magistral enregistré en situation, un entretien avec deux interlocuteurs, une réunion associative avec de nombreux locuteurs.

Un premier volet du travail a consisté à évaluer les fonctionnalités des plateformes : sécurité et confidentialité des données, tarification, interopérabilité, simplicité d'utilisation, outil d'édition[9]. Un second volet a consisté à évaluer les transcriptions obtenues. La méthodologie employée pour comparer les transcriptions est innovante. 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 mesurée du potentiel de gain de temps de transcription pour chacun des fichiers et des plateformes. Les volets 2 et 3 sont novateurs par rapport aux travaux existants.

Notre démarche montre que deux outils se démarquent : Vocapia et Sonix, chacun ayant ses domaines de prédilection. Ensuite, suivent d'assez près Go Transcribe et Happy Scribe qui d'après nos observations semblent partager des technologies très similaires. Video indexer et YouTube sont à considérer pour leur coût modique et leurs performances pour certains corpus. Headliner semble un peu retrait, mais se distingue en proposant des fonctionnalités annexes de montage et d'incrustation de sous-titres. De plus, c'est le seul avec YouTube qui propose des heures de transcriptions gratuites. Vocalmatic présente des performances similaires à Headliner, mais sans les fonctions annexes et la gratuité d'utilisation. *In fine*, aucune plateforme ne serait plus efficace que les autres pour l'ensemble des extraits audio. La qualité de la transcription dépend du type de fichier soumis en entrée : certaines seront plus performantes sur des discours planifiés, d'autres sur la parole spontanée. En revanche, toutes échouent à retranscrire le fichier le plus complexe, la réunion associative, de manière exploitable en raison de la présence de nombreux chevauchements, extraits inaudibles et bruits de fond. Quel que soit le type de fichier ou de plateforme, un temps de réécoute et de correction reste indispensable, pour un gain de temps final observé pouvant aller jusqu'à 75 % par rapport à une transcription manuelle.

L'utilisation de ces outils s'accompagne d'un certain nombre de questionnements liés à la confidentialité et la sécurité des données, quelles que soient leur nature : techniques (ex. adresse IP, version du navigateur), nominatives (ex. téléphone, coordonnées bancaires), et qu'elles soient issues de corpus audio ou de texte transcrit. L'usage fait par les plateformes de ces données, notamment, reste obscur. L'utilisation de ces outils n'est donc pas recommandée dans le cadre de données sensibles ou confidentielles.

Enfin, indépendamment de la qualité de ces outils, le principe même de leur utilisation peut être soumis à critique. D'une part, parce que l'étape de transcription est considérée par de nombreux chercheurs comme une étape de l'analyse, qui ne saurait être déléguée à une machine. D'autre part, car l'automatisation d'un certain nombre de métiers n'est pas sans poser des questions sociétales, à la fois en matière de pertes d'emplois et de conditions de travail des opérateurs humains qui équipent parfois ces dispositifs. Sans oublier le travail caché que peuvent représenter les heures d'édition en ligne fournies gratuitement par les utilisateurs et les utilisatrices et qui permettent d'améliorer les algorithmes de ces outils.

Mots-clés

Transcription automatique – Corpus oraux – Retranscription entretien – Évaluation logiciels – Méthodologie – Données de la recherche

[9] Ces aspects ont été étudiés à la date d'élaboration de ce rapport (mai 2020) et ont pu évoluer.

Remerciements

Nous tenons tout d'abord à remercier le Comité MATE-SHS, sans qui cette aventure n'aurait jamais vu le jour.

Nous remercions également le Comité d'organisation des Tuto@MATE pour son invitation à présenter une première version de ce travail, et tous les participants et participantes de la session pour leur intérêt, questions et commentaires.

Merci à Matthieu Quignard (I.C.A.R.) pour sa relecture attentive sur les métriques de distance, et à Karine Onfroy (GREThA) pour ses retours critiques.

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 enfin 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 mener à bien ce projet.

Ressources associées

Ce travail a été présenté dans le cadre des Tuto@MATE en juin 2020.

La vidéo de ce webinaire est disponible à l'adresse https://mate-shs.cnrs.fr/actions/tutomate/tuto24-retranscription-elise-tancoigne/ ainsi que les transparents diffusés.

Table des matières

I. Origine et fonctionnement du projet

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 rappeler brièvement.

Ce projet a pour particularité d'avoir démarré de façon spontanée suite à une discussion sur la liste de diffusion du réseau méthodologique MATE-SHS[10], porté par des ingénieurs et des ingénieures du CNRS et outillant la recherche en Sciences Humaines et Sociales (SHS). Ce réseau national compte environ 500 membres. En mai 2019, une abonnée s'enquiert sur la liste d'avis et conseils concernant Vocalmatic, une plateforme automatique de transcription de fichiers audio. Aucun membre n'était alors en mesure de lui répondre, mais plusieurs personnes se sont manifestées pour mentionner des expériences avec d'autres plateformes. L'une d'entre nous a alors proposé de mutualiser les connaissances existantes :

« Je me demandais si cela intéresserait quelques personnes de faire un benchmark de ces nouveaux outils : on pourrait choisir un entretien commun (libre d'utilisation et dans l'idéal déjà retranscrit), se répartir les plateformes, et comparer ensuite les résultats obtenus. » (Élise Tancoigne, 14 mai 2019)

Très rapidement, une quarantaine de personnes s'est déclarée intéressée par le sujet, dont environ la moitié était prête à participer aux tests. La discussion a alors continué hors liste. Il s'agissait au départ de se répartir le travail pour compléter un tableur collectivement. Mais 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. La première réunion a eu lieu en visioconférence deux mois plus tard et comptait 6 personnes de profils différents : sociologie, linguistique, informatique, ingénierie audiovisuelle, sciences politiques, géographie. Elle fut suivie d'environ 15 nouvelles rencontres virtuelles, d'abord au rythme de tous les deux mois, puis tous les mois et enfin toutes les semaines sur la fin du projet. Le petit groupe initial a perduré. Certaines personnes ont également contribué de manière ponctuelle et sont donc mentionnées comme contributrices au rapport.

Nous avons au départ travaillé en définissant collectivement les tâches à effectuer et en répartissant le travail par plateformes, ce qui a permis à chacun d'avoir une vision d'ensemble de la chaîne de transcription[11]. Nous avons ensuite adopté une approche nous permettant d'avoir un regard transversal sur tous ces outils. Nous nous sommes réparti des tâches d'analyse par sous-groupes qui ont travaillé chacun de leur côté sur toutes les plateformes. Un premier groupe a réalisé l'état de l'art sur les outils d'évaluation des résultats de transcription automatique de la parole, et procédé à des calculs de proximité entre textes. Un second groupe a procédé à une évaluation fine de la nature des erreurs trouvées dans les textes retranscrits. Un troisième groupe s'est attelé à une évaluation qualitative des transcriptions obtenues. Enfin, le travail de

[10] https://mate-shs.cnrs.fr/

[11] Nous avons privilégié l'emploi d'infrastructures et d'outils de collaboration institutionnels ou libres et gratuits : réseau Renater , Jitsi Meet, Framacalc, UTBox (Université de Tours), Sharedocs (Huma-Num).

comparaison des fonctionnalités des plateformes, qui avait initialement été réparti entre tous les membres du groupe, a également été repris par deux d'entre nous.

Ce projet a donc la spécificité d'avoir été réalisé entièrement à distance dès le départ, sans budget, en collaboration entre des personnes qui ne se connaissaient absolument pas et portaient chacune un intérêt différent à ces outils. Et, surtout, qui n'avaient absolument pas pris la mesure du travail que cela représenterait au final… À ce jour, les membres ne se sont toujours pas rencontrés « in real life » comme disent les aficionados de jeux vidéo… Ironie de l'histoire, c'est Vocalmatic, la plateforme à l'origine de notre démarche, qui remporte les moins bons résultats parmi tous les outils testés.

II. Qu'a-t-on cherché à évaluer ?

Contexte

Le recueil de la parole à travers des entretiens de divers types (individuel ou en groupe, libre ou dirigé) 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 et un peu décourageante. Pour 1 heure d'enregistrement, la durée de transcription peut en effet s'étendre de 4 à 6 h (Rioufreyt, 2016 : 11), voire 30 h (Lamberterie et al., 2006) ou plus selon l'expérience de l'opérateur, les caractéristiques de l'enregistrement et de la transcription (convention de transcription, granularité, phénomènes pris en compte, annotation des balises temporelles…).

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. En effet, 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*, le « *gain de performance important obtenu récemment grâce aux méthodes de DeepLearning et aux réseaux de neurones profonds* », et la « *création de technologies de plus en plus robustes aux 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 ». 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, soit uniquement basées sur la similarité lexicale des sorties obtenues, soit sur les tarifications. Parmi les comparatifs les plus élaborés que nous ayons trouvés, ceux de Tim Bunce (2018, 2019, 2020) font partie des plus aboutis. Cela dit, bien que ce dernier ait intégré à ses comparaisons des plateformes de service de transcription « humaine » ou mixte, ce qui nécessitait un budget non négligeable, et qu'il ait raffiné de façon intéressante la préparation de ses fichiers tests, la métrique sur laquelle il a établi son « palmarès » final reste un unique indicateur statistique basé sur la proximité lexicale (le WER : Word Error Rate), que nous avons utilisé également, mais pas uniquement. Il apparaît ainsi que le test comparatif mis en œuvre ici par la complémentarité de ses approches (quantitatif et qualitatif) constitue une proposition originale. En outre, il n'existait à ce jour aucune comparaison pour la langue française. Ce travail est donc la première évaluation approfondie de ces plateformes pour un usage sur des enregistrements en français.

Posture

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)[12], les outils de dictée vocale (ex. Dragon, VoiceNote)[13], 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).

Nous n'aborderons pas ici tous les usages et développements liés à la reconnaissance vocale et sa transcription pour des applications technologiques industrielles et commerciales. Dans le seul domaine de la recherche, la variété des disciplines en SHS se traduit par la variété des éléments que chacune va rechercher lors de l'utilisation des plateformes de transcription. Claire-Blanche Benveniste écrit à ce sujet que « 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 improvisés. 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 donc avant tout :

- à faciliter la préparation de leur corpus en vue d'une analyse du discours (dans le sens « contenu du discours »). Cela se caractérise par le souhait d'accéder à des verbatims de qualité (avec éventuellement les « heu », les hésitations, les répétitions), avec une bonne identification des locuteurs (le travail de découpage pouvant être très important, que ce soit dans un entretien en tête-à-tête ou un entretien collectif) ;
- à générer automatiquement du sous-titrage de vidéo et de l'indexation de contenu vidéo : une attention particulière est alors portée à la possibilité de notation des bornes chronologiques des énoncés, ou « timecode » ;
- à effectuer une analyse conversationnelle ou une analyse de la production linguistique : on souhaite alors une transcription la plus proche possible du signal, sans ponctuation, mais avec l'intégralité de ce qui a été prononcé (onomatopées, amorces et répétitions, chevauchements, pauses, notation de l'intonation…).

Le Tableau II-1 permet d'illustrer la traduction concrète de ces différences dans les transcriptions recherchées.

[12] Pour une comparaison de ces outils, consulter Rioufreyt (2018).
[13] Nous avons testé un certain nombre d'entre eux (Dragon V15, Google docs, VoiceNote, Dictée vocale Mac et Windows), mais ils ne sont pas adéquats pour travailler sur des fichiers préenregistrés. Soit ils ne permettent pas le chargement d'un fichier externe, soit ils perdent rapidement pied : un travail de ralentissement des enregistrements serait nécessaire pour pouvoir les utiliser.

Entretien verbatim	Sous-titrage	Analyse conversationnelle	Analyse de la production linguistique
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	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	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:pper avoir au PRP:det au tel NOM:trc téléphone tout à l'heure ADV 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

Tableau II-1 Un même extrait transcrit pour différents usages

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 chercheurs :

a. Fonctionnalité des plateformes
   - Sécurité et confidentialité des données
   - Tarification des services de transcription
   - Caractéristiques des transcriptions et formats (XML, DOC, etc.)
   - Facilité de prise en main de l'interface
   - Qualité des outils d'édition disponibles sur la plateforme pour retravailler le texte
b. Segmentation du texte
   - Respect de l'alternance des locuteurs
   - Possibilité d'avoir des balises temporelles fines
c. Précision lexicale
   - Précision absolue du discours : respect des interjections/onomatopées, marques du travail de formulation : « heu », « hein », hésitations, répétitions, faux départs, auto-réparations, disfluences, etc.
   - Précision du discours sans ces marques du travail de formulation
   - Respect des règles orthographiques, de syntaxe et d'accord
   - 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.

Nous avons donc mené deux grands types d'évaluations : celle des fonctionnalités et caractéristiques des plateformes d'une part, et celle des transcriptions automatiques obtenues, d'autre part.

Trois approches complémentaires

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. Les résultats sont présentés dans la section IV.

L'évaluation des résultats a quant à elle été déclinée selon trois approches :

1. Une approche de comparaison classique par la métrique du WER (section V)
2. Une approche qualitative permettant de caractériser et comprendre les erreurs (section VI)
3. Une évaluation du gain de temps obtenu par rapport à une transcription manuelle (section VII).

Ces trois approches se complètent mutuellement. Le WER permet d'avoir un premier classement des plateformes, mais ne permet pas de comprendre ce qui se joue derrière ce classement. Sur quoi butent-elles ? 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*) ? La deuxième 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. La troisième approche est 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 ?

III. Corpus de référence

Choix des corpus

Le choix d'un corpus de référence a constitué une phase importante de nos réflexions initiales. En effet, ce corpus devait correspondre à un cahier des charges qui comportait plusieurs points.

Nous souhaitions tout d'abord que les corpus retenus soient **représentatifs des données que nous utilisons dans le cadre de nos activités de recherche** ainsi que du travail d'une large communauté de chercheurs en SHS. Nous visions notamment des disciplines comme la sociologie, la linguistique, la psychologie et la didactique. Nous avons donc recherché :

- une hétérogénéité de situations interactionnelles (discours académique, conversation professionnelle ou institutionnelle, entretien de recherche, conversation familière…),
- une variété de phénomènes linguistiques propres aux interactions verbales (chevauchements, hésitations, amorces, réparations, onomatopées…),
- une variété dans le nombre de locuteurs,
- une variété dans la qualité audio des extraits.

Ensuite, ces corpus devaient être **compatibles avec les contraintes résultant de notre démarche méthodologique**. Celle-ci nécessitant des temps d'analyse qualitative importants, elle a en partie déterminé les caractéristiques du corpus en matière de nombre d'échantillons retenus, de durée des extraits et de complexité intrinsèque. Rappelons ici l'absence de budget propre, qui a également influé sur le choix du corpus de référence.

Nous avons également pris en compte **la capacité du corpus à faire l'objet d'une transcription automatique**. En effet, la littérature ainsi que l'expérience acquise dans le cadre de nos activités de recherche montrent que tous les corpus ne sont pas adaptés à ce type de traitement. Il en va ainsi des conversations comprenant une densité de chevauchements importante, un bruit de fond prononcé, des problèmes liés à la qualité d'enregistrement, etc. Cependant, afin d'évaluer les limitations des plateformes étudiées dans le traitement de ces phénomènes, il nous semblait important que notre corpus intègre au moins un extrait qui comporte ces types de phénomènes.

Une fois ces critères définis, il nous a fallu déterminer la provenance des corpus. Pour des raisons pratiques, nous avons opté pour l'utilisation de corpus issus de nos travaux de recherches antérieurs. Ce choix présentait au moins deux avantages, la maîtrise des contenus et l'assurance qu'ils répondaient aux règles déontologiques et légales à respecter dans le cadre d'une étude scientifique. Sur la base de ces critères, nous avons retenu dans un premier temps une petite dizaine d'extraits audio. Puis nous avons rapidement écarté la moitié d'entre eux, qui nous semblaient relever d'un genre discursif trop spécifique ou difficilement exploitable. Pour finir, nous avons retenu quatre extraits audio, tous au **format Mpeg 1/2 Layer 3 (Mp3)**. En voici les caractéristiques :

`Physionomie` : le sous-corpus `Physionomie` correspond à un texte littéraire extrait des essais de Montaigne. Le texte, en version française modernisée, est lu par un narrateur professionnel dont la diction peut être qualifiée de parfaite. 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 en français modernisé) dans des conditions de diction optimales.

`Camille` : le sous-corpus `Camille` correspond à un extrait d'entretien scientifique entre deux locuteurs, dont l'un connaît des difficultés d'énonciation dues à un handicap physique. 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 la parole de locuteurs rencontrant des difficultés phonatoires ou une voix atypique.

`Comptines` : l'extrait `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 ici d'évaluer la faculté des outils à restituer un monologue académique en public similaire à ceux produits dans le cadre d'une conférence, d'un discours politique, d'un cours magistral, etc.

`Harmonie` : cet extrait est tiré d'une réunion associative. Il comporte une dizaine de locuteurs et les conditions d'enregistrement ne sont pas très bonnes (présences de bruits ambiants, locuteurs éloignés du dispositif de captation…). Il se caractérise par un grand nombre de chevauchements et la présence de conversations parallèles parfois chuchotées. Cet extrait est représentatif d'une situation de conversation spontanée de groupe dans un contexte associatif ou professionnel.

Une fois le corpus sélectionné, nous devions déterminer la durée des extraits. Au départ, une durée de 20 minutes pour chacun d'entre eux était envisagée, dans l'idée que les algorithmes seraient peut-être plus performants sur de longs extraits. Mais compte tenu des moyens humains dont nous disposions, combiné aux conditions d'accès aux plateformes payantes dont les temps d'essais sont généralement limités à 30 minutes, **nous avons finalement opté pour des échantillons d'une durée de 5 minutes**.

Cette durée peut a priori paraître courte. Cependant, elle permet d'intégrer une variété considérable de phénomènes dans un contexte de production orale. De plus, la densité verbale de nos extraits, qui comptent entre 660 et 800 mots transcrits et des temps de pause réduits, nous laissait à penser qu'ils seraient adaptés aux besoins de notre étude. En outre, cette durée se situe dans la moyenne des études préexistantes, qui privilégient généralement des échantillons de 4 à 8 minutes.

Le Tableau III-1 synthétise les descriptions des corpus retenus. Chacun de ces corpus cristallise une ou plusieurs difficultés potentielles pour les plateformes.

Physionomie	Comptines	Camille	Harmonie
Texte lu	**Cours universitaire en présentiel**	**Entretien sociologique**	**Réunion associative**
Excellente qualité audio	**Bonne qualité audio**	**Bonne qualité audio**	**Bruits ambiants**
Vocabulaire ancien	**Monologue académique en public**	**Difficultés phonatoires**	**Nombreux locuteurs**
Projet MONLOE[14]	**Projet TCOF[15]**	**© Laure Gayraud**	**Projet TCOF**
5 min — MP3 — 660 à 800 mots			

Tableau III-1 Caractéristique des extraits audio retenus pour les tests

Création de transcriptions de référence

Nous dénommons **transcriptions originales** les textes provenant de transcripteurs différents et d'origines disparates et qui ne peuvent constituer, en l'état, une référence. Il nous est en effet apparu le besoin d'une **transcription de référence** pour comparer les mots de la transcription avec l'audio. Celle-ci n'est pas à considérer comme exacte, puisque les tests montrent que la transcription humaine comporte aussi une variabilité, des choix et des habitudes divers (Bilger, 2008). Cependant, nous devions nous accorder sur une transcription dont la séquence des mots pouvait être comparée à celle de la transcription automatique fournie par chaque plateforme testée. Ce travail préparatoire a été réalisé par un membre du groupe. Chaque extrait audio a été réécouté et les choix suivants ont été faits :

- suppression des noms de locuteurs et des balises temporelles (ces codes lorsqu'ils étaient présents n'étaient pas harmonisés entre les sous-corpus)
- suppression des codes et annotations phénomènes (les plateformes ne produisent pas d'annotation dans les sorties au format texte)
- conservation et harmonisation des *heu* (euh, heu) ainsi que des répétitions
- conservation de la ponctuation
- transformation des chiffres écrits en toutes lettres en chiffres (ex. *9 heures 30*)
- harmonisation des apostrophes
- suppression des espaces surnuméraires

Cette première harmonisation a donné lieu à ce que l'on a appelé les **transcriptions manuelles**. Le membre du groupe a ensuite adopté une convention de transcription qu'il a appliquée à tous les textes fournis avec les extraits audio (pour un récit plus détaillé de la procédure, voir **Annexe** 1), après réécoute. Les fichiers obtenus ont constitué les **transcriptions de référence**. Un exemple en est donné dans le Tableau III-2.

[14] Disponible sur https://montaigne.univ-tours.fr/category/multimedia/ed-sonore/
[15] Disponible sur https://tcof.atilf.fr/ et https://hdl.handle.net/11403/tcof/v2

Avant harmonisation (Transcription originale)	Après harmonisation (Transcription de référence – REF)
vous avez, il y a une date de limite je pense, pour euh accÈder au bureau quand on est au C.A. il y a une… euh on va dire une < date oui, oui. par rapport > ‡ l'A.G. oui il faut il faut faire partie de de la ah non < non absolument pas, tu peux Ítre du C.A. ou non > vous n'avez pas Áa dans les statuts, < parce que dans certains je pense pas > statuts il doit y avoir au moins six mois	vous avez, il y a une date de limite je pense, pour accéder au bureau quand on est au C.A. il y a une… on va dire une date oui, oui. par rapport à l'A.G. oui il faut il faut faire partie de de la ah non non absolument pas, tu peux être du C.A. ou non vous n'avez pas ça dans les statuts, parce que dans certains je pense pas statuts il doit y avoir au moins six mois

Tableau III-2 Exemple d'harmonisation d'une transcription de référence

Chaque transcription de référence a ensuite été enregistrée dans le même format (texte brut UTF8, avec saut de ligne Windows).

Harmonisation des transcriptions obtenues

Les mêmes normes d'harmonisation ont été appliquées aux **sorties brutes** obtenues par les plateformes. Celles-ci peuvent en effet suivre des conventions différentes qui font que la transcription sera considérée comme fausse ou « en erreur » par rapport au référentiel adopté, alors que le sens est correct.

Quelques exemples de « normes de transcription » correctes, mais variables et donc susceptibles de générer de « fausses erreurs » et d'empêcher la comparaison entre les résultats :

- la transcription des nombres : 1 000, 1000 ou mille.
- les noms composés : plate-forme ou plateforme
- autres normalisations : XXIème siècle, 21è s. etc. ou certaines abréviations comme OGM, AOC.

Après harmonisation, les fichiers ont là encore été enregistrés en format texte (.TXT), UTF8, avec saut de ligne Windows.

IV. Comparaison des fonctionnalités des plateformes

Nous nous sommes rapidement aperçus qu'une comparaison basée uniquement sur des indicateurs de performance était insuffisante pour rendre compte de la pertinence et de l'intérêt des outils. Quels seraient par exemple, les bénéfices d'utiliser une plateforme aussi performante soit-elle, si par exemple, la confidentialité et la sécurité des données ne sont pas garanties ou si elle est incompatible avec les formats de données que nous exploitons dans le cadre de nos activités de recherche ?

C'est pourquoi nous avons établi une liste de critères en phase avec les besoins et exigences inhérents à nos disciplines, nous permettant ainsi de comparer de la manière la plus objective possible l'ensemble des outils.

Nous avons identifié cinq grandes thématiques, chacune regroupant une partie des critères étudiés. La première concerne la sécurité des données. Elle englobe la sécurité physique des données (protocoles d'échanges, caractéristiques de l'hébergement…), mais également les aspects juridiques liés à la consultation, l'exploitation et l'archivage des données. La seconde porte sur la tarification des services. La troisième est dédiée aux caractéristiques des transcriptions restituées, en termes de fonctionnalités et présence de métadonnées (balises temporelles, étiquette locuteur…). La quatrième partie aborde la question des formats acceptés et restitués tant pour les corpus que pour les transcriptions. Enfin la cinquième et dernière partie est consacrée à une évaluation de la simplicité d'utilisation de l'interface ainsi qu'à la richesse fonctionnelle de l'éditeur de transcription.

Ces cinq volets sont détaillés ci-dessous. À noter qu'au sein de ce chapitre nous utiliserons indistinctement les termes plateformes, outils, logiciels ou sites pour désigner une même entité, la plateforme étudiée.

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. Il en va ainsi des informations nominatives rattachées au compte (nom, prénom, adresse, mail, coordonnées bancaires…), mais également des informations techniques envoyées lors de la navigation sur les sites (adresse IP, version du navigateur, informations de connexion, cookies, géolocalisation…). La collecte et le croisement de ces informations à l'aide de technologies de type « data mining », peuvent conduire les éditeurs à élaborer des profils commerciaux des utilisateurs dans une optique d'exploitation marketing. Les corpus eux-mêmes pourraient, techniquement, faire l'objet d'analyse à des fins différentes de celles qui sont annoncées, par exemple l'exploitation des données dans un but commercial ou industriel.

Pour ces raisons, il a paru important de vérifier que la sécurité et la confidentialité des données étaient garanties. Pour ce faire nous avons examiné plusieurs éléments décrits et synthétisés dans les paragraphes suivants.

Critères étudiés

Présence de texte définissant les conditions d'utilisation des services (CGU, mentions légales…)

Tous les services comportent des CGU (Conditions Générales d'Utilisation) ou équivalents, plus ou moins détaillés (Tableau IV-1). On constate que les textes décrivent de manière relativement exhaustive les conditions d'utilisation du site. Néanmoins, ils sont moins explicites quant à la manière dont les données sont utilisées. À l'heure actuelle, la plupart des mentions légales sont rédigées en anglais (à l'exception de YouTube, disponible également en français).

	Happy Scribe	Headliner	Go Transcribe	Sonix	Video Indexer	Vocalmatic	Vocapia	YouTube
Présence CGU ou mentions légales (O/N)	O	O	O	O	O	O	O	O
Anglais/ Français	Anglais	Anglais	Anglais	Anglais	Anglais	Anglais	Anglais	Français

Tableau IV-1 Présence ou absence de Conditions Générales d'Utilisation

Conditions d'accès aux données personnelles (conformité au Règlement Général sur la Protection des Données)

L'utilisation de ces outils est conditionnée à l'obtention d'un compte d'utilisateur 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 l'exploitation qui est faite de ces données conformément aux règles édictées par le Règlement Général sur la Protection des Données (RGPD). Rappelons qu'en principe, le RGPD qui est une directive qui émane de l'U.E., 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'U.E. et ce, quel que soit le pays où les données sont hébergées. Il s'en suit pour les États concernés, une superposition et une cohabitation des règlements régissant l'utilisation des données numériques, pas foncièrement triviales à implémenter du côté des entreprises et à interpréter par les utilisateurs du service. Cela dit, les plateformes étudiées qui possèdent leur siège social hors de la CEE sont censées respecter le RGPD du fait qu'elles échangent et exploitent des données soumises par des clients européens.

Précisions de nature juridique

La souscription d'un service de transcription est un contrat entre le souscripteur et le fournisseur du service de transcription.

Il conviendra dans un premier temps de distinguer qui est le souscripteur, personne physique (étudiant, doctorant, chercheur) ou personne morale (ex. laboratoire, université). L'attribution des responsabilités en dépend et s'avère précieuse lorsqu'un problème dans l'exécution du contrat surgit. Si c'est une personne morale, il est conseillé d'avoir sollicité et recueilli l'avis favorable du service juridique afférent s'il existe.

Le droit français précise que les documents nécessaires au consentement du souscripteur doivent lui permettre de garantir que son consentement n'est pas entaché ; par conséquent ils doivent être rédigés dans la langue du souscripteur et mis à disposition par le fournisseur du service.

Celui ou celle qui souhaite souscrire un service de transcription est en droit et légitime à solliciter auprès du fournisseur de service, non seulement les CGU écrites en français, mais également tous les autres documents relatifs à la fourniture du service de transcription.

Sur la question de la protection des données, l'U.E. s'est dotée d'un dispositif, le *Bouclier de Protection des Données*

Le *Bouclier de Protection des Données* repose sur des engagements pris par des sociétés américaines pour respecter les principes, les règles et les obligations fixés par le dispositif du Bouclier de Protection des Données UE-États-Unis :

— le principe d'information

— le principe du choix, selon lequel les personnes doivent pouvoir effectuer certains choix s'agissant de la transmission de leurs données à des tiers, de l'utilisation de leurs données pour des finalités substantiellement différentes ou en cas de traitement de données sensibles

— le principe d'intégrité des données et de finalité du traitement

— l'obligation d'assurer la sécurité des données

— l'obligation de protéger les données lorsque celles-ci sont transférées à une société tierce.

Aussi, ce dispositif accorde un certain nombre de droits aux personnes dont les données à caractère personnel ont été transférées d'une entité européenne vers les États-Unis :

— le droit d'être informé d'un tel transfert,

— le droit d'exercer ses droits d'accès, de rectification et de suppression des données à caractère personnel qui ont été transférées.

À savoir : Il est possible de vérifier si une société établie aux États-Unis adhère au dispositif du Bouclier de Protection des Données en consultant la liste disponible en ligne sur www.privacyshield.gov. Pour en savoir plus : Guide relatif au Bouclier de Protection des Données UE — Etats-Unis (Commission Européenne, 2019).

Concernant l'accès aux données, tous les sites consacrent dans leurs mentions légales un chapitre dédié à la consultation, la modification, la rectification et la suppression de ces données (Tableau IV-2).

	Go Transcribe	Happy Scribe	Headliner	Sonix	Video Indexer	Vocalmatic	Vocapia	YouTube
Nom de la société	Go-Transcribe Ltd	Happy Scribe Ltd	SpareMin	Sonix Inc.	Microsoft	Enactics Inc.	Recherche 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-territorialité)	O (extra-territorialité)	O (extra-territorialité)	O (extra-territorialité)	O	O (extra-territorialité)
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 Frameworks

Tableau IV-2 Règlementations appliquées. N/D = non déterminé

Nous constatons à travers ces mentions que dans la majorité des cas le droit de disposer de leurs données est ainsi reconnu aux utilisateurs. Cependant les procédures permettant d'effectuer ces opérations sont décrites de manière plus ou moins détaillée selon les prestataires. À noter que certains sites se réservent le droit de facturer des frais liés à la recherche des informations demandées.

Remarques concernant les données des tableaux présentés dans ce chapitre :

1) les informations recueillies sur les différents prestataires sont issues des textes présents sur leur site Web et réseaux sociaux exploités par l'entreprise (Facebook, Twitter…). Nous avons porté une attention particulière aux Conditions Générales d'Utilisation. Cependant le déchiffrage de ces dernières ne constitue pas une tâche aisée pour les non-juristes que nous sommes. 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. Compte tenu de cette part d'incertitude, nous ne pouvons qu'encourager les utilisateurs de ces services à bien considérer la sensibilité de leurs données avant de les soumettre. En effet, nous n'avons aucune certitude quant à la façon dont elles sont exploitées, mettant parfois à mal les clauses de confidentialités qui peuvent accompagner certains corpus.

2) Certains critères que nous ne sommes pas en mesure de certifier n'ont pas été reportés sur le tableau. Il en va ainsi du lieu d'hébergement des données ou de la déclaration CNIL. Lorsqu'une valeur associée à un critère manque, nous la notons « non déterminée » (N/D).

3) Les données présentées sont celles en vigueur au moment de l'étude (mai 2020). Elles sont donc susceptibles de varier dans le temps.

Confidentialité et exploitation des données nominatives

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 (à noter que nous nous interrogeons sur la portée et les conséquences de ces partages). Par ailleurs, quelques sites comme Sonix ou Vocalmatic proposent une inscription en utilisant les données associées à d'autres services (Facebook, 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 données par ces entités tierces.

Exploitation des données techniques

Tous les sites étudiés recueillent un certain nombre de données techniques (adresse IP, version du navigateur, caractéristiques du système d'exploitation, cookies…). À l'exception de Headliner, ces sites stipulent que les données nominatives ne sont utilisées qu'à des fins administratives ou de service. Il convient là aussi d'être prudent pour les raisons qui seront exposées dans le paragraphe suivant.

Exploitation des corpus

La plupart des mentions légales stipulent que les données du corpus sont uniquement utilisées à des fins d'amélioration des modèles et technologies de reconnaissance. Seulement, il convient d'être circonspect concernant ces propos, pour diverses raisons.

Tout d'abord, car la plupart des sites étudiés utilisent une ou plusieurs technologies externes pour le traitement des données. Certains (en minorité) le revendiquent clairement, d'autres le font de manière plus opaque ou n'en font tout simplement pas mention. Nous constatons en effet dans le cadre de nos analyses 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. Ainsi Headliner ou Vocalmatic exploitent clairement cette ressource. D'autres comme Sonix, Go Transcribe et Happy Scribe semblent également entretenir un lien avec lui sans que nous puissions l'expliciter. 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 localisé aux États-Unis (E.U.). 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 et par conséquent impose aux sociétés de conserver une copie de toutes les données pendant un temps variable. Le Cloud Act s'applique aussi bien aux sociétés qui possèdent leur siège social aux E.U. qu'aux sociétés étrangères qui hébergent des données dans ce pays ainsi qu'aux sociétés américaines qui hébergent des données dans le reste du monde. C'est le cas par exemple de YouTube, Google Speech, Video Indexer. Ce qui implique également qu'une société européenne, qui sous-traiterait l'hébergement de ces données à une société américaine, serait soumise aux mêmes

contraintes juridiques que cette dernière. C'est pourquoi il convient d'être extrêmement prudent et prudente vis-à-vis des données déposées sur ces sites, surtout si elles comportent un caractère confidentiel ou sensible.

Protocoles de cryptage des données échangées

L'ensemble des plateformes utilisent des protocoles sécurisés pour assurer les échanges de données avec leurs infrastructures qui 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 (Tableau IV-3). Ce qui en pratique signifie qu'il y a peu de chances que les données soient interceptées durant une phase d'envoi ou de récupération des données. Par ailleurs, la plupart des sites mettent en avant l'argument de la sécurité des données tout au long de la chaîne de traitement et de stockage. Selon elles, tout est fait pour la garantir à travers un certain nombre de procédures (redondance des stockages, chiffrement des données entreposées, sauvegardes régulières, vérification antivirus, utilisation de pare-feu…). Mais le propos est ensuite nuancé en indiquant que les prestataires ne peuvent pas être tenus responsables en cas de problèmes de perte ou de corruption. Par conséquent, nous vous conseillons de toujours conserver une copie des données sur vos machines ou supports de stockage externes.

	Go Transcribe	Happy Scribe	Headliner	Sonix	Video Indexer	Vocalmatic	Vocapia	YouTube
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

Tableau IV-3 Protocoles de cryptage

Communication avec le prestataire

Nous avons recensé les modes de contact avec les prestataires (Tableau IV-4). Tous les services en proposent au moins un. Ils se déclinent sous la forme de formulaires, de modules de discussion en direct (chat), d'adresse mail de contact, de forums… Il convient ici de préciser que nous n'avons pas vérifié la portée et l'efficacité de ces services.

	Go Transcribe	Happy Scribe	Headliner	Sonix	Video Indexer	Vocalmatic	Vocapia	YouTube
Hotline (O/N)	O	O	O	O	O	O	O	N
Moyens de contact	Mail, Téléphone selon abonnement	Chat, Mail	Chat, Blog, FAQ, Formulaire	Mail, Formulaire, Téléphone selon abonnement	Formulaire, Forum, Téléphone	Chat	Téléphone, Mail, Formulaire	FAQ, Forums

Tableau IV-4 Communication avec le prestataire

Tarification des services de transcription

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, selon sa complexité et l'expérience du transcripteur, une durée minimale de travail de 4 à 6 heures pour une heure de signal. Ce qui, selon la nature de la transcription, représente un coût *a minima* de 60 et 100 euros pour une heure de signal[16]. Certains sites 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 (Tableau IV-5). 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 (cf. chapitres sur l'analyse des performances des plateformes) vis-à-vis du type de corpus que l'on souhaite retranscrire afin de ne pas engendrer des coûts inutiles. Nous pouvons conseiller deux stratégies pour cela : la première consiste à déposer un échantillon du corpus sur le site puis en fonction des résultats, traiter la totalité du corpus. La seconde consiste à utiliser un site concurrent gratuit pour effectuer des essais sur un échantillon.

Enfin, notons que la plupart des plateformes proposent un système de parrainage qui permet d'obtenir des heures de transcription gratuites.

	Go Transcribe	Happy Scribe	Headliner	Sonix	Video Indexer	Vocalmatic	Vocapia	YouTube
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 académique	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

[16] 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 %

	Go Transcribe	Happy Scribe	Headliner	Sonix	Video Indexer	Vocalmatic	Vocapia	YouTube
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 personnalisée	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 filigrane), 350 Mo max. par fichier	N (30 mn d'essai)	O (10 heures de média ; 40 heures à travers l'API[17])	N (30 mn d'essai)	N (essai sur demande)	O Plateforme gratuite
Système de parrainage (O/N)	O	O	O	O	N	O	N	N/A

Tableau IV-5 Tarification des services de transcription au 31/05/2020
N/A = non adapté ; N/D = non déterminé

Caractéristiques et métadonnées des transcriptions

Il s'agit d'évaluer ici la capacité des logiciels à associer certaines données secondaires ou annotations en lien avec les transcriptions (Tableau IV-6). Il en va ainsi de la génération des balises temporelles, de la présence d'un champ permettant de désigner le locuteur, de la détection automatique des changements de locuteurs avec création d'un nouveau tour de parole, de l'identification automatique des locuteurs, de la transcription de la ponctuation et enfin de la possibilité d'enrichir le lexique. Quelques-uns de ces aspects sont abordés dans le chapitre dédié à l'éditeur de transcription. Cependant, rappelons succinctement à quoi ils font référence.

- *Langues supportées* : pour des questions de temps et de moyens, notre corpus comportait uniquement des textes en français. Cependant, nous avons reporté dans notre tableau de synthèse les langues supportées. Cette liste repose uniquement sur les déclarations des prestataires.
- *Génération des balises temporelles (timecode)*. Il s'agit ici de savoir si l'outil est capable de restituer les balises temporelles qui encadrent un tour de parole, indépendamment du fait qu'on puisse les modifier en ligne ou pas. En général, l'unité de temps utilisée est le millième de seconde.
- *La détection automatique des changements de locuteurs* : nous cherchons à vérifier si l'outil étudié opère un saut de ligne lorsqu'un changement de locuteur se produit. Dans le cadre de cette étude, nous n'avons pas évalué l'efficacité de la fonction.
- *L'identification automatique des locuteurs* : l'outil est-il capable d'identifier automatiquement les différents locuteurs qui participent à une conversation, et de leur attribuer un pseudonyme unique de locuteur (ex. Loc1, Loc2…) ? L'efficacité de cette détection automatique n'est pas évaluée ici.
- *La présence d'un champ permettant de désigner le locuteur*. Dans le cas d'une transcription, il est souvent intéressant de pouvoir désigner le locuteur par

[17] Portion de code informatique permettant d'utiliser les plateformes sans passer par l'interface Web

son nom ou (le plus souvent) par un pseudonyme. Nous vérifions qu'un champ dédié à cette identification manuelle est proposé.

- *La transcription de la ponctuation :* il s'agit de vérifier si l'outil transcrit les signes de ponctuation. Signalons que lorsque cette fonction est proposée, nous n'avons pas vérifié la pertinence de cette ponctuation.
- *La possibilité d'enrichir le lexique :* il s'agit de vérifier si l'outil autorise l'ajout de termes à son lexique (entités nommées, termes spécialisés, techniques…). Nous n'avons pas cherché à vérifier si l'enrichissement du lexique permettait d'améliorer la performance de la transcription.
- *La transcription des marques de formulation et disfluences*. Nous cherchons à savoir si les outils restituent les onomatopées (*heu, hein, bah, ben, hum…), les répétitions, les bégaiements*…

	Go Transcribe	Happy Scribe	Headliner	Sonix	Video Indexer	Vocalmatic	Vocapia	YouTube
Langues supportées	Principales langues occidentales et slaves. Japonais	Principales langues occidentale, arabes, asiatiques et slaves	Principales langues occidentale, arabes, asiatiques et slaves	Principales langues occidentale, arabes, asiatiques et slaves	Principales langues occidentale, arabes, asiatiques et slaves	Principales langues occidentale, arabes, asiatiques et slaves	Principales langues occidentale, arabes, asiatiques et slaves,hébreu, hindi, pachto, swahili, ourdou	Allemand, anglais, russe, coréen, italien, espagnol, français, japonais, néerlandais, 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
Enrichis-sement du lexique (O/N)	N	O (vocabulaire spécifique) 100 termes max.	N	O (vocabulaire spécifique)	O (personnalisation des modèles)	N	O (personnalisation des modèles sur demande)	N
Marques d'élocution (O/N)	N	N	N	N	Partielle	N	Partielle	Partielle

Tableau IV-6 Caractéristiques et métadonnées des transcriptions. N/D = non déterminé

Formats de fichiers en entrée et en sortie

Ce chapitre est dédié aux formats d'import/export, dont dépendent les traitements effectués et l'exploitation des sorties. Nous distinguons deux types de formats. Ceux rattachés aux données déposées, c'est-à-dire les fichiers son et vidéo destinés à être retranscrits (formats d'entrée ou d'import) et ceux restitués par les outils qui correspondent au résultat de la transcription (formats de sortie ou d'export).

Formats en entrée

À l'exception de YouTube qui n'accepte que des fichiers vidéo[18], toutes les plateformes testées sont en mesure de charger des fichiers audio encapsulés dans les formats les plus courants : le Mpeg 1/2 Layer 3 (MP3) qui est un format compressé avec perte d'informations, et le Waveform Audio File (WAV), format sans perte d'information (Tableau IV-7). Le format FLAC (format compressé sans perte) est également accepté par certaines plateformes. Les fichiers de type AIFF (format historique utilisé par Mac OS) ne sont disponibles que sur Sonix, Vocapia et Happy Scribe.

En ce qui concerne les formats filmiques, les fichiers MPeg4 ou Quicktime (.MOV) sont reconnus par l'ensemble des outils à l'exception de Vocapia qui ne gère pas les vidéos. Les fichiers .AVI sont acceptés sur tous les sites sauf ceux de Vocalmatic et de Headliner. Le tableau suivant indique les formats reconnus ainsi que la taille maximum acceptée pour les fichiers.

	Go Transcribe	Happy Scribe	Headliner	Sonix	Video Indexer	Vocalmatic	Vocapia	YouTube
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 transcription 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)

Tableau IV-7 Formats de fichiers en entrée.
N/D = non déterminé

[18] Si l'on souhaite utiliser les fonctions de transcription à partir d'un fichier audio, il faut préalablement le convertir en fichier vidéo. L'opération peut se faire à l'aide d'un logiciel de montage vidéo en ajoutant une piste vidéo issue d'un autre fichier. On obtient ainsi un fichier dont la vidéo et l'audio n'ont pas de lien mais qu'il sera possible de déposer sur la plateforme.

Formats en sortie

Les formats de fichiers proposés par les outils pour récupérer les transcriptions sont très variés (Tableau IV-8). Les utilisations envisagées de ces transcriptions peuvent dans certains cas conditionner le choix d'une plateforme. Certaines sont plutôt dédiées à la génération de sous-titres de vidéo comme Video Indexer ou YouTube qui restituent des fichiers Subrip (.SRT), WebVTT (.VTT), SubViewer (.SBV). Un exemple de fichier VTT est donné dans le Tableau II-1 page 5 (colonne 2). Cependant en général, les outils permettent d'exporter les transcriptions dans des formats compatibles avec les traitements ou éditeurs de texte (.DOC, .TXT). C'est le cas pour Go Transcribe, Sonix, Happy Scribe, Video Indexer, Vocalmatic et Vocapia. Remarquons que seule la plateforme Vocapia propose un export en XML qui intègre l'indexation des balises temporelles à l'échelle du mot ainsi qu'un indice de confiance pour la transcription. Pour sa part, Headliner est l'unique prestataire qui propose en option une incrustation de la transcription directement dans le fichier vidéo.

	Go Transcribe	Happy Scribe	Headliner	Sonix	Video Indexer	Vocalmatic	Vocapia	YouTube
Formats transcription texte en sortie	DOC, TXT, PDF, SRT, VTT	DOC, PDF, HTML, TXT, VTT, SRT, STL, Première	VTT, MP3, MP4 (incrustation transcription)	DOC, TXT, SRT, VTT FCPxml, xmlPremière	SRT, VTT, TTML, TXT, CSV	DOC, TXT	XML, VTT, RTF, PDF	SRT, VTT, SBV

Tableau IV-8 Formats de fichiers en sortie

Remarques concernant les fichiers de sous-titre (SRT, VTT, SBV…)

La plupart des plateformes (à l'exception de Vocalmatic) proposent de restituer les transcriptions au format SRT, VTT ou SBV. Ces formats sont conçus pour permettre la synchronisation des sous-titres avec les vidéos auxquelles ils sont associés et sont généralement utilisés par des logiciels de lecture vidéo (VLC…)[19]. Par conséquent, ils ne sont pas adaptés à une exploitation directe dans un traitement de texte à moins de prévoir une étape de conversion ou de formatage.

En résumé, les plateformes sont compatibles avec les formats multimédias les plus courants. Il faut distinguer les formats liés aux données multimédias et ceux en relation avec les transcriptions. Pour l'import des données, la plupart des outils acceptent les formats audio courants comme le WAV et le MP3. Concernant les vidéos, ce sont les formats compatibles avec la norme MPEG 4 qui sont privilégiés. Au niveau de l'export et à l'exception notable de Headliner qui permet de restituer le résultat des montages opérés sur son site et d'incruster la transcription dans le film, seuls les formats basés sur du texte sont restitués (DOC, TXT, SRT…), ce qui en pratique signifie que le résultat de la transcription sera exporté sous la forme d'un document textuel séparé. Vocapia quant à lui, permet d'exporter un fichier XML annoté comportant des informations utiles dans certains contextes. Elles comportent notamment les segments vocaux et non vocaux, les étiquettes des locuteurs, les balises temporelles à l'échelle du mot, des scores de confiance et les signes de ponctuation.

[19] La plupart des lecteurs multimédias affichent ces sous-titres en surimpression de la vidéo à condition que 1) les fichiers vidéo et de sous-titres aient le même nom (mais leur propre extension) et 2) qu'ils soient placés dans le même dossier que le film associé.

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

Tous les services intègrent un éditeur permettant de relire et corriger en ligne le résultat des transcriptions. La présence d'un tel éditeur participe à l'efficacité et à l'intérêt global de la plateforme.

De même, la plupart des outils permettent d'effectuer un alignement entre l'audio ou la vidéo et la transcription correspondante (la transcription défile en même temps que l'enregistrement associé). Cette opération est rendue possible par la présence de balises temporelles affectées aux énoncés lors de la transcription automatique. Cette fonction de synchronisation entre les enregistrements et les énoncés est utile à de nombreux chercheurs. Elle permet de se référer à un segment visuel ou auditif à partir d'un énoncé et vice-versa.

Cependant, cet alignement n'est pas aisé à entreprendre en utilisant des logiciels bureautiques (traitement de texte…) conjointement à des lecteurs/éditeurs multimédias comme Audacity ou VLC. En effet, les nombreux allers-retours entre les logiciels pour effectuer le report des repères de temps constituent des tâches chronophages et potentiellement génératrices d'erreurs. Les logiciels de type CAQDAS (logiciel d'aide à l'analyse de données qualitatives) comme Elan, Nvivo, Sonal, Transana, intègrent des fonctions permettant d'effectuer ce travail. Seulement, leurs coûts d'entrée en termes monétaire comme de formation n'en font pas non plus une solution idéale pour réaliser ce type de tâche. De plus, aucun d'entre eux ne permet la mise à jour automatique des balises temporelles lors d'une segmentation de la phrase et il n'existe aucun logiciel grand public proposant cette fonction.

Au-delà des questions de segmentation temporelle, nous avons vérifié la présence de certaines fonctions qui nous semblaient importantes comme la possibilité d'associer un tour de parole à un locuteur, la mise en évidence des termes susceptibles d'être mal transcrits (score/indicateur de proximité par rapport à l'enregistrement…). Les autres critères examinés sont liés aux formats d'importation et d'exportation de la transcription éditée ainsi qu'aux fonctions d'aide à la transcription.

Critères étudiés

Présence d'un éditeur de texte en ligne

Toutes les plateformes intègrent un éditeur plus ou moins fonctionnel. Une grande majorité d'entre elles permettent la modification des balises temporelles et certaines proposent un champ modifiable dédié au nom du locuteur (Tableau IV-9).

	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	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

Tableau IV-9 Éditeur de transcription : description.
N/A = non adapté ; N/D = non déterminé

Alignement temporel de la transcription sur le signal

L'ensemble des plateformes propose l'affichage et l'édition des balises temporelles associés aux transcriptions (Tableau IV-10). Cependant, seuls Go Transcribe, Sonix et Happy Scribe intègrent une fonction de segmentation des énoncés au niveau du mot. Elle s'accompagne d'un réalignement automatique entre la transcription et les balises temporelles associées. Les autres plateformes (hormis Vocapia) permettent de modifier la segmentation temporelle manuellement.

Nous avons également vérifié la présence d'une fonction permettant d'ajuster le « timing » de départ de la transcription (par défaut réglée sur 00:00:00.000).

	Go Transcribe	Happy Scribe	Headliner	Sonix	Video Indexer	Vocalmatic	Vocapia	YouTube
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

Tableau IV-10 Éditeur de transcription : balises temporelles.
N/A = non adapté ; N/D = non déterminé

Mise en forme des caractères/paragraphes

À la manière d'un logiciel de traitement de texte, quelques outils autorisent la mise en forme des caractères (gras, italique…) ainsi que des paragraphes (alignement, retraits…). Nous considérons cette possibilité comme étant moins stratégique que la précédente dans le sens où c'est une opération simple à mener à l'aide d'un logiciel dédié. Néanmoins, la chose peut présenter un intérêt non négligeable lorsqu'on souhaite appliquer aux fichiers de sous-titres une mise en forme des caractères (taille de la police, couleur…).

Ces enrichissements sont réservés aux formats compatibles avec les logiciels de traitement de texte (DOC, RTF…) et dans une moindre mesure aux fichiers de sous-titres (SRT, VTT). Les formats qui reposent sur du texte brut (TXT) ou le CSV en sont donc exclus.

Aide à la transcription

Les transcriptions automatiques nécessitent des corrections plus ou moins importantes pour être exploitées. Dans ce contexte des fonctions d'assistance à la transcription sont les bienvenues. La plupart sont liées à l'écoute et à la navigation dans l'extrait audio (réglage de la vitesse d'écoute, écoute du signal avec « rembobinage » de quelques secondes, affichage de la forme d'onde, raccourcis clavier pour la lecture…). D'autres sont liées à l'interface (mise en forme du texte, fluidité de l'interface, 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 GoTranscibe. Les moins efficaces dans ces opérations sont Vocalmatic et Vocapia.

	Go Transcribe	Happy Scribe	Headliner	Sonix	Video Indexer	Vocalmatic	Vocapia	YouTube
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

Tableau IV-11 Éditeur de transcription : aide à la transcription

Synthèse

	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
Modification manuelle segmentation des balises temporelles (O/N)	O	O	O	O	N	N	N	O
Réalignement automatique des balises temporelles lors d'une segmentation (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
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'exportation 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, TTML, TXT, CSV	Word, TXT	XML, Word, TXT,	VTT, SRT, SBV
Notation globale de l'éditeur (1 à 5)	**4**	**4,5**	**3,5**	**4,5**	**3**	**3**	**2,5**	**3,5**

Tableau IV-12 Éditeur de transcription : tableau de synthèse.
N/A = non adapté ; N/D = non déterminé

La présence d'un éditeur facilite la correction des transcriptions et participe à l'intérêt de ces outils. Nous nous sommes particulièrement attachés à vérifier la présence de fonctions liées à la correction des transcriptions ainsi que l'ergonomie générale de l'éditeur : présentation, fluidité et stabilité, simplicité d'utilisation. L'examen de ces

points nous a permis d'attribuer une note globale[20] sur une échelle de 1 (la plus faible) à 5 (la plus élevée) (Tableau IV-12).

Nous constatons que la qualité et les fonctionnalités des éditeurs sont très variables. Au moment où nous rédigeons ce rapport, ceux de Sonix, Go Transcribe et Happy Scribe sont très complets et bien adaptés aux besoins des transcripteurs. Celui de Vocapia est vraiment en retrait du point de vue de la richesse fonctionnelle. Entre temps, le site YobiYoba.com qui émane de la société Vocapia a été lancé en 2020[21]. Il partage les mêmes modèles de reconnaissance automatique que ce dernier, tout en proposant une interface et un éditeur d'une grande qualité.

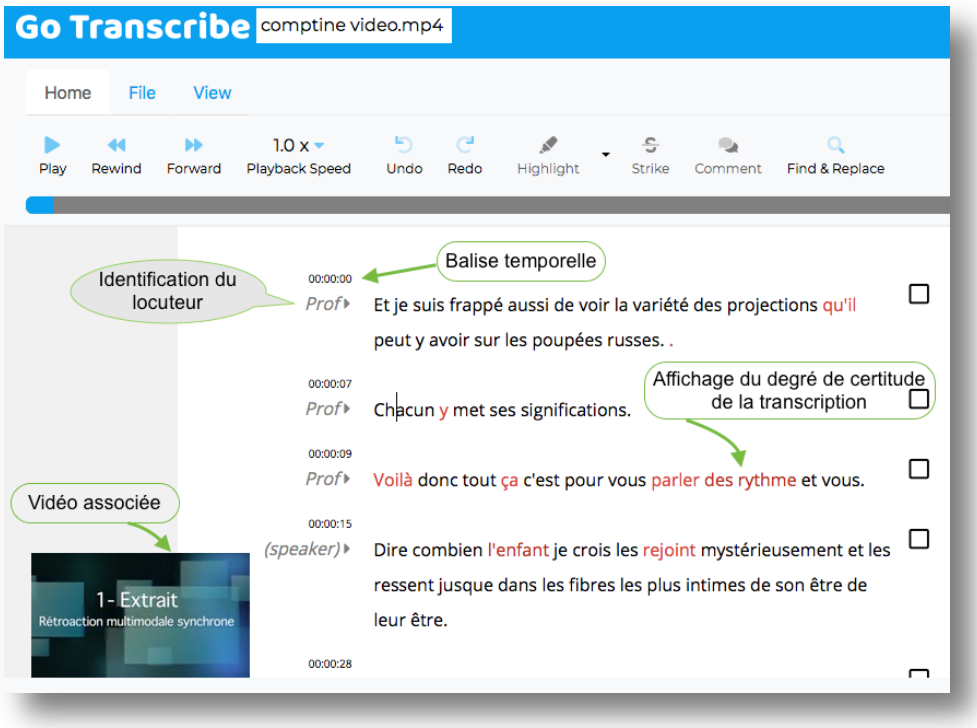
Copies d'écran commentées

Ce paragraphe propose un descriptif succinct de chaque outil et éditeur associé. Les copies d'écran sont assorties d'un descriptif de l'éditeur et d'un commentaire. Pour les outils qui ne sont pas spécifiquement conçus pour la transcription automatique ou qui comportent des fonctions spécifiques, une rubrique «remarques» décrivant ses particularités a pu être ajoutée.

Go Transcribe

4/5

L'éditeur est simple d'emploi. Il intègre toutes les fonctions utiles à la correction d'une transcription et des métadonnées associées. Des champs sont dédiés au nommage des locuteurs et à la modification des balises temporelles. La segmentation d'un énoncé met automatiquement à jour les balises temporelles associées. La navigation dans le signal est aisée.



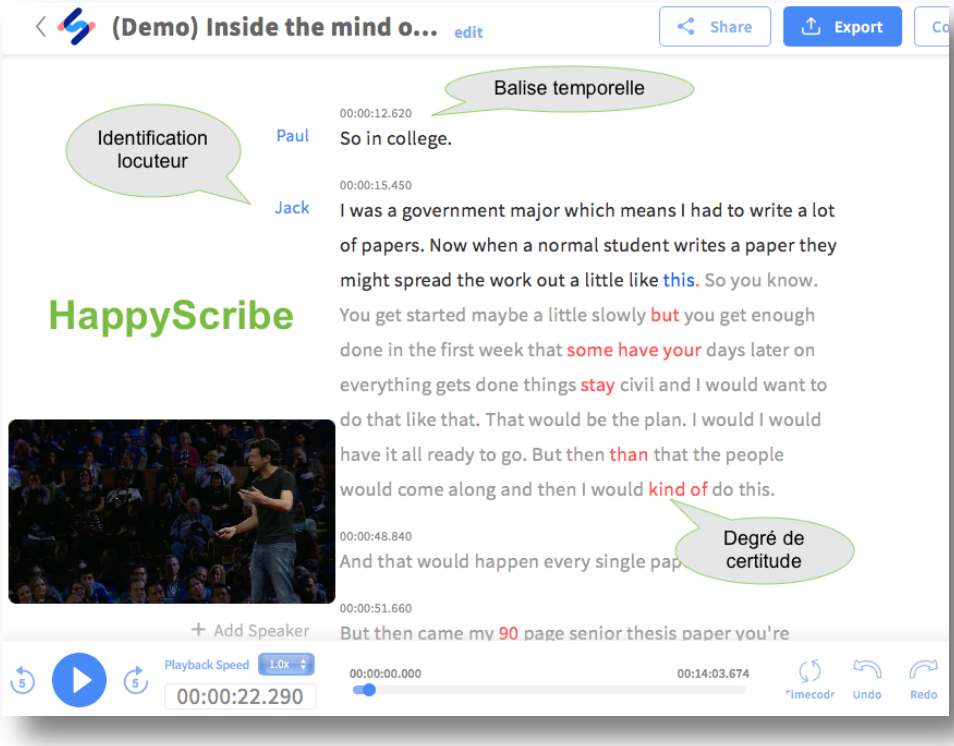
[20] Cette note est calculée sur la base de la richesse fonctionnelle de l'outil et de sa facilité d'utilisation. Elle a été attribuée par deux membres du groupe rompus à l'utilisation d'une large gamme de logiciels dans le domaine du traitement et de l'analyse de données plurimédias. Nous insistons 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.

[21] La société Vocapia invite les chercheurs à utiliser YobiYoba pour les corpus d'une durée inférieure à 24h.

Happy Scribe

4,5/5

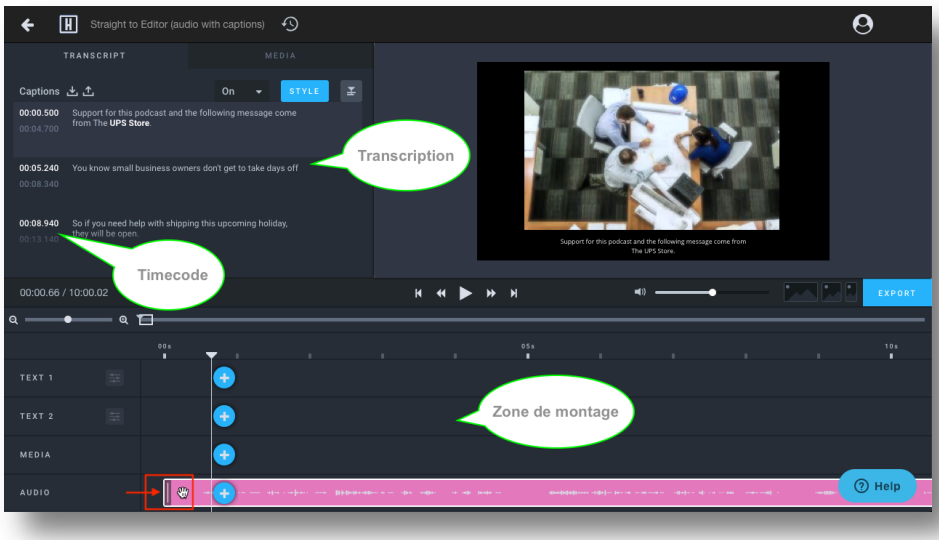
Happy Scribe profite d'une interface fonctionnelle. Les balises temporelles et les étiquettes des locuteurs sont facilement repérables et modifiables. La segmentation des lignes s'accompagne d'un réajustement automatique des balises temporelles. Une mini interface de navigation dans le signal est proposée et des fonctions d'aide à la transcription sont présentes.



Headliner

3,5/5

Headliner est un logiciel de montage audiovisuel en ligne qui propose des fonctions de transcription automatique. L'interface est assez bien conçue dans le cadre d'un montage vidéo. C'est le seul outil qui permet d'exporter les films en incrustant les transcriptions dans la vidéo. Nous regrettons cependant l'absence d'étiquettes dédiées aux noms des locuteurs. De plus la segmentation des énoncés et la correction des balises temporelles n'est pas très aisée et s'effectue par couper-coller.

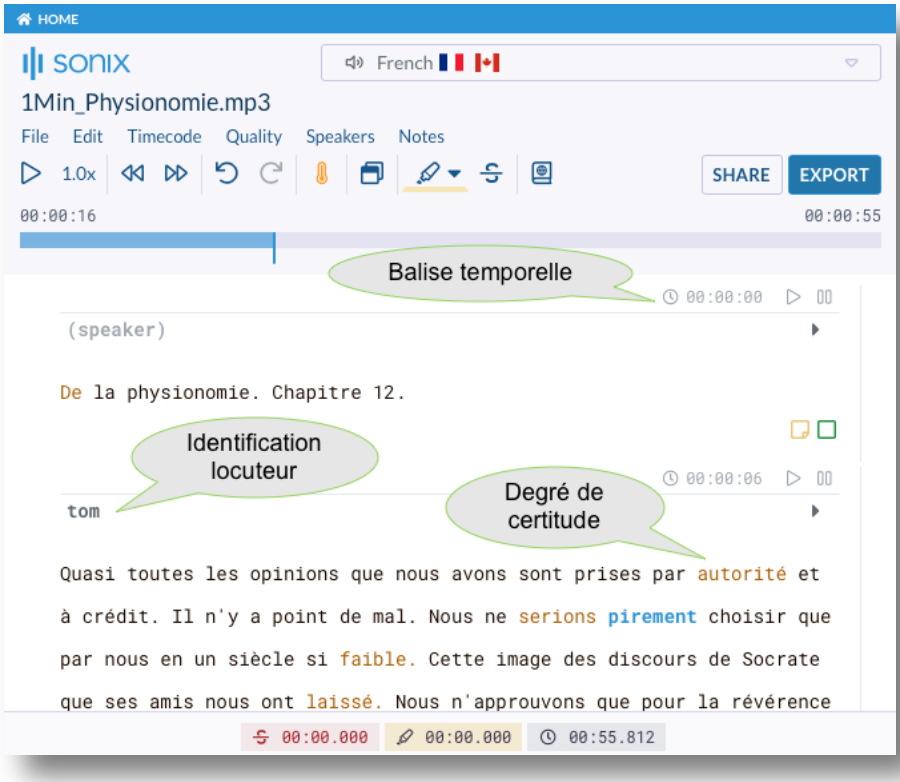


Remarques : Headliner propose également d'autres fonctions comme la création de fichiers audio avec forme d'onde animée, l'incrustation de sous-titres… Il peut être utilisé à travers une API dédiée facilitant ainsi son utilisation à travers un script.

Sonix

4,5/5

L'éditeur se révèle riche en fonctionnalités et facile d'accès grâce à une présentation proche d'un logiciel de traitement de texte. Les balises temporelles et les étiquettes locuteur sont bien conçues et facilement modifiables. Le degré de certitude permet de mettre le focus sur les segments incertains et le réalignement des balises temporelles fonctionne parfaitement. La navigation dans le signal est facilitée par la présence d'un mini lecteur multimédia. Enfin, l'export des fichiers bénéficie également d'une large palette de formats.

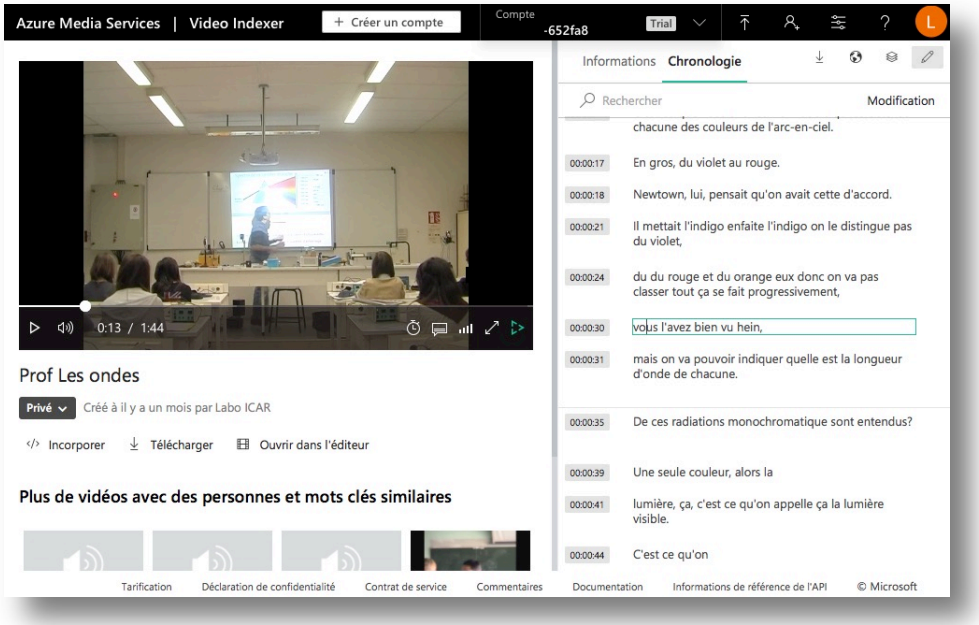


Remarques : Sonix propose de nombreuses autres fonctionnalités selon le type d'abonnement. Nous pouvons par exemple citer les fonctions collaboratives qui se caractérisent par une gestion fine des accès accordés aux utilisateurs et aux fichiers. Il est également possible d'importer une transcription existante afin de générer les balises temporelles correspondantes.

Video Indexer

3/5

La présentation de l'éditeur est bien conçue pour un usage orienté vers l'indexation de vidéos. En revanche les étiquettes des locuteurs et des balises temporelles sont difficilement modifiables et la segmentation des énoncés complexe (il faut effectuer des couper-coller pour passer un fragment de texte à la ligne).

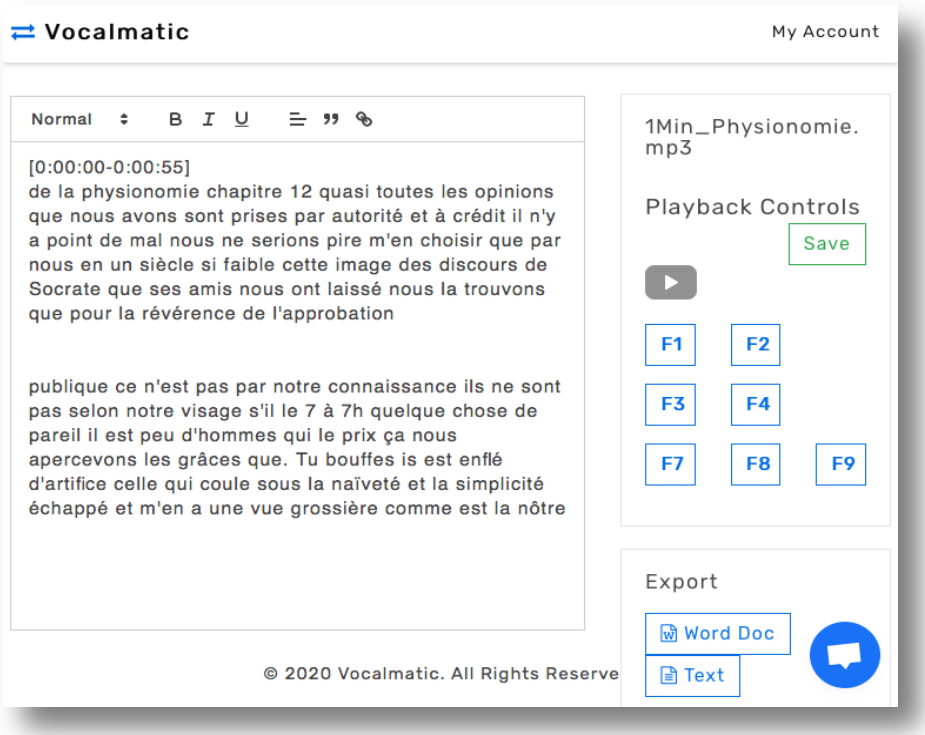


Remarques : Video Indexer est conçu pour l'indexation des vidéos (détection automatique de scènes, des plans, fonctions de réduction du bruit, production de statistiques de temps de parole de l'orateur, détection des silences, des applaudissements…). La transcription automatique est possible, mais ce n'est clairement pas la fonction principale de l'outil. Il propose également une API facilitant son intégration et son utilisation à travers un script.

Vocalmatic

3/5

Cet outil propose une interface épurée, mais fonctionnelle. Nous regrettons l'absence des étiquettes locuteurs ainsi que d'une fonction de réalignement automatique des balises temporelles. Les formats d'export sont également limités (Word et texte brut). Des fonctions d'aide à la transcription sont présentes.



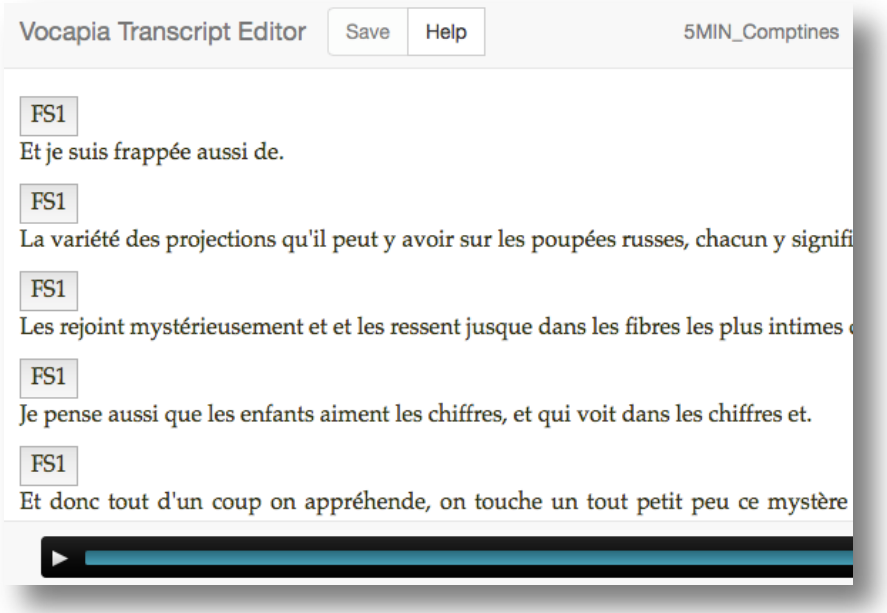
Vocapia

2,5/5

L'interface de l'éditeur est assez décevante. Les fonctions sont limitées et l'affichage des balises temporelles est absent.

Cependant nous apprécions la compatibilité de l'interface avec la plupart des navigateurs, l'alignement précis des mots par rapport au signal et les touches de fonction pour écouter le son.

Une détection automatique des locuteurs est possible, mais elle se révèle encore peu précise.

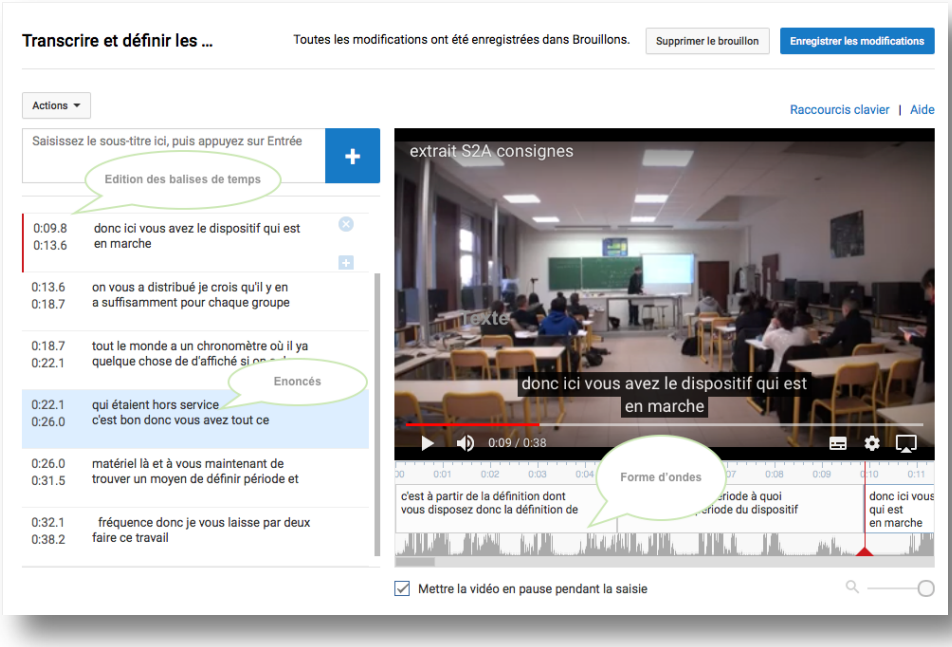


Remarques : Vocapia est une société française de R&D qui travaille en partenariat avec le LIMSI (CNRS), laboratoire qui mène des recherches sur le traitement des

langues. Il se distingue entre autres des autres outils par l'existence d'une suite logicielle « VoxSima » pour Linux. Pour les gros corpus, une API est disponible pour un coût d'un centime d'euro la minute (60 cts de l'heure) qui autorise son utilisation à travers un script. Vocapia repose sur ses propres modèles de transcription automatique. Il est le seul à proposer une sortie XML intégrant des segments vocaux et non vocaux, des étiquettes de locuteur, des codes temporels annotés à l'échelle du mot, des scores de confiance et des signes de ponctuation. Enfin, la plateforme accepte un traitement des fichiers par lots et depuis peu, les technologies Vocapia sont accessibles pour le traitement de petits corpus (< 24 h) à travers le site filial YobiYoba.com.

YouTube

3,5/5



YouTube profite d'une interface riche. Elle se caractérise par des fonctions de montage vidéo permettant par exemple de segmenter les films ou de flouter les visages. Elle offre également une fonction intéressante de synchronisation entre le signal et des transcriptions préalablement importées. Cependant certains éléments de l'interface manquent d'intuitivité. De plus, les étiquettes de locuteurs sont absentes et les exports se font uniquement dans des formats conçus pour le sous-titrage (SRT, VTT, SBV).

Remarques : Le site d'hébergement de vidéos YouTube propose une fonction de transcription automatique dans le cadre de son service en ligne. Pour pouvoir en bénéficier, il faut préalablement créer un compte sur le service ou utiliser un compte Google existant. Ensuite, il suffit de créer une « chaîne » pour accueillir vos vidéos (les fichiers audio ne sont pas acceptés) et définir la visibilité de ces dernières (publique ou privée). A moins que vous ne souhaitiez diffuser largement vos contenus, nous vous recommandons de les rendre privés.

Fonctionnalités additionnelles

Nous avons examiné la présence de fonctions supplémentaires qui nous paraissaient utiles dans le cadre d'une recherche (Tableau IV-13) :

1) Présence d'une API : Une API est une portion de code informatique qui permet d'utiliser les plateformes sans passer par l'interface Web. Ce script peut faire l'objet d'une intégration dans un script plus large stocké sur votre machine ou diffusé en ligne. Il permet généralement d'étendre les possibilités de traitement

de l'outil (par exemple, en automatisant l'envoi des fichiers, en convertissant automatiquement les formats…).

2) Le traitement par lot : il s'agit ici de vérifier si l'outil est capable de transcrire plusieurs fichiers en même temps à travers l'interface en ligne.
3) Fonctions collaboratives : certains outils autorisent le partage de comptes afin de mettre en place une répartition des tâches ou une organisation du travail (relectures, corrections, ajouts...) au sein d'une équipe.

	Go Transcribe	Happy Scribe	Headliner	Sonix	Video Indexer	Vocalmatic	Vocapia	YouTube
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 (sur service en ligne)	N
Fonctions collaboratives (O/N)	N	O	N	O (selon abonnement)	O	N	O	O

Tableau IV-13 Fonctionnalités supplémentaires

N/D = non déterminé

Ce panorama, à la fois synthétique et suffisamment détaillé a été conçu pour se projeter concrètement dans l'utilisation des outils en matière de sécurité des données, de tarifs, de caractéristiques des transcriptions restituées, d'interopérabilité et de fonctionnalités associées, en particulier l'éditeur. À ce stade, les utilisateurs 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 ».

V. Comparaison par calcul de distance

Ce chapitre est consacré aux indicateurs d'évaluation quantitative automatique de la proximité de deux textes. Un état de l'art expose ce que serait une mesure d'évaluation idéale puis présente les métriques disponibles, leur intérêt et leurs limites. Il s'appuie largement sur le travail de revue d'Errattahi et al. (2018) et la thèse de Ben Jannet (2015)[22].

Selon McCowan et al. (2004), une mesure d'évaluation idéale de la reconnaissance automatique de la parole devrait être : (i) **directe**, c'est-à-dire mesurer la composante reconnaissance automatique de la parole indépendamment de ses applications finales, (ii) **objective**, c'est-à-dire calculée de manière automatisée, (iii) **interprétable** : la valeur absolue de la mesure doit donner une idée de la performance, et (iv) **modulaire** : la mesure d'évaluation doit être générale, pour permettre une analyse approfondie en fonction de l'application.

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[23], en travaillant au niveau du mot plutôt qu'au niveau du caractère. Autrement dit, il 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

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

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.***				

[22] La thèse de Ben Jannet est orientée sur la reconnaissance des entités nommées. Dans le but d'évaluer ses propres propositions, ce dernier a étudié en détail les outils d'évaluation des résultats obtenus dans le domaine de la reconnaissance automatique de la parole et leur a consacré d'importants développements.
[23] 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.

Le WER se calculera de la manière suivante :

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

Limites du WER

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 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 effet, s'il peut y avoir un nombre de suppressions limité, dépendant du nombre de mots initial, le nombre d'insertions est, quant à lui, illimité. De sorte que dans des conditions de bruit, le WER peut dépasser 100 %, en donnant beaucoup plus de poids aux insertions qu'aux suppressions (Errattahi et al., 2018). Ben Jannet (2015) constate ainsi que « [dans le cas de] l'intégration des systèmes de RAP dans des chaînes de traitement où d'autres modules utilisent leur sortie comme entrée, de nombreuses études ont remarqué des incohérences entre les mesures données par le WER et les performances obtenues au niveau de l'application globale ».

La critique essentielle du WER est qu'il est efficace pour la reconnaissance vocale où les erreurs peuvent être corrigées par la frappe, comme par exemple la dictée, mais qu'il ne fournit aucun détail sur la nature des erreurs. Or certaines erreurs sont plus perturbatrices que d'autres, certaines pouvant être corrigées plus facilement que d'autres (comme par exemple des erreurs de flexion : *échappé vs échappées*).

Enfin, si l'objectif n'est pas une transcription parfaite mot à mot, mais la précision de la compréhension globale du langage, le WER n'est pas la meilleure métrique. Il existe en effet une divergence entre le taux d'erreur des mots et cette compréhension : 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 %.

Des mesures alternatives ont ainsi été proposées pour résoudre ces limites du WER.

Alternatives au WER

Information mutuelle

Plutôt que d'estimer le coût de restauration d'une séquence de mots originale (ce que fait le WER), certains chercheurs ont tenté d'estimer la proportion d'information communiquée par la transcription obtenue. La première mesure créée dans ce sens était le **Relative Information Lost (RIL)**, ou « Information Relative Perdue » qui a été proposée dès les années 50 par Miller (1955) puis reprise par Woodard et Nelson (1982), Maier (2002) et Morris et al. (2004).

L'information mutuelle (IM, ou MI) fournit ici une mesure de la dépendance statistique entre les mots d'entrée X et les mots de sortie Y. En théorie de l'information, elle mesure la quantité d'information apportée en moyenne par une réalisation de X sur les

probabilités de réalisation de Y : H(Y | X), H étant le nombre de mots communs entre la séquence de départ et la séquence obtenue. Les différentes probabilités qu'elle requiert pour son calcul peuvent être estimées à partir des fréquences relatives obtenues dans une matrice de confusion des mots en entrée et en sortie.

Néanmoins, le RIL n'a pas été vraiment utilisé : non seulement parce qu'il n'est pas simple à appliquer, mais également parce qu'il présente des limites dans la prise en compte des erreurs. Ben Jannet (2015) confirme et précise : « mis à part le fait que le RIL est une métrique assez lourde à implémenter (notamment l'estimation de H(Y | X)), il ne permet pas de prendre en compte les erreurs systématiques (qui se répètent toujours de la même façon) […]. Ce cas peut être particulièrement accentué pour les mots peu fréquents (dans le test), pouvant être des mots hors vocabulaire et potentiellement des entités nommées (informations riches sémantiquement). Le RIL est également une métrique très dépendante de l'alignement[24] comme le montre l'étude de Maier (2002), dans laquelle l'auteur teste plusieurs algorithmes d'alignement et montre que les mesures données par le RIL changent significativement avec le type d'alignement ».

La seconde mesure créée, appelée **Word Information Lost** (**WIL**), est une mesure d'approximation du RIL, plus simple à appliquer, car basée uniquement sur les comptages utilisés pour le WER. Proposée par Morris et al. (2004), elle est donnée sous la forme :

$$WIL = 1 - \frac{\mathrm{H}^2}{(\mathrm{H} + \text{sub.} + \text{supr.})(\mathrm{H} + \text{sub.} + \text{aj.})}$$

H représente le nombre de mots communs entre la séquence de départ et la séquence obtenue. Les expériences menées par (McCowan et al., 2004 ; Morris et al., 2004) ont montré que le RIL et le WIL peuvent être des mesures intéressantes lorsque les systèmes de RAP possèdent des taux d'erreur élevés.

Ainsi Morris et al. (2004) démontrent que les termes WER et WIL ne sont pas toujours classés de la même façon. Ben Jannet (2015) observe également des résultats « non stables » dans sa thèse. Pour l'un de ses deux jeux de données tests, « *le WIL et la F-mesure ne sont pas de meilleures mesures que le WER dans le contexte applicatif étudié* ».

Autres mesures

Nanjo et al. (2005) ont défini une mesure d'évaluation de compréhension de la parole permettant de pondérer les erreurs obtenues : le taux d'erreur pondéré sur les mots-clés (**Weighted Keyword Error Rate, WKER**). Ils ont principalement travaillé sur des corpus de discours planifiés (cours, présentations orales). Cette mesure donne plus de poids à certains mots, considérés comme plus importants dans le discours (les mots-clés).

[24] L'alignement s'entend ici comme la mise en relation des segments textuels de deux corpus. Ainsi, l'alignement des deux énoncés *Jean mange* et *J'en mange* consiste à mettre en relation *Jean* et *J'en*, aussi bien que *mange* et *mange*.

Cependant l'étude de Park et al. (2008) n'a pas permis d'identifier une différence significative entre l'utilisation du WER et du WKER.

Favre et al. (2013) ont proposé une mesure d'évaluation alternative au WER, le **H-score**, en partant du principe que la mesure ne devait pas seulement porter sur la qualité de la transcription en référence à un signal d'origine, mais également en termes d'adéquation avec les utilisations postérieures qui peuvent en être faites : recherche d'information, compréhension orale, traduction automatique. Leur corpus comprenait des transcriptions de réunions d'entreprise devant être utilisées a posteriori pour faire un audit des décisions managériales réalisées. Cependant la solution proposée, valide, pose des problèmes importants de mise en œuvre. Elle repose sur des algorithmes classificateurs qui prennent en compte l'incorporation de jugements humains sur les transcriptions obtenues. Or ces jugements sont longs et coûteux à obtenir, bien loin des objectifs que nous nous sommes fixés dans ce travail.

Posture adoptée

Errattahi et al. (2018) concluent leur examen des méthodes de détection et de correction des erreurs dans les systèmes de transcription par le constat qu'à ce jour, il est nécessaire d'approfondir les recherches sur le sujet, et qu'il convient d'accorder une attention particulière à des questions telles que l'efficacité, la facilité d'utilisation et la robustesse des méthodes développées.

Le WER est une mesure simple et largement utilisée pour évaluer la précision de la transcription et les alternatives ne semblent pas encore parfaitement crédibles. Ben Jannet (2015), dans la conclusion de sa thèse portant sur la recherche de nouveaux indicateurs (pour les entités nommées) confirme ce point : « Des alternatives au WER ont été proposées dans la littérature. Nous distinguons particulièrement RIL (Relative Information Loss) et WIL (Word Information Lost) qui sont fondées sur la mesure de la perte d'informations […]. Toutefois, ces métriques n'ont pas été très utilisées ni testées sur des données réelles pour prouver leur efficacité. »

Ainsi, malgré ses lacunes, « le WER est la métrique classiquement utilisée pour évaluer la qualité des sorties de RAP. Cette métrique a prouvé son efficacité quand il s'agit d'évaluer les systèmes de RAP en isolation [c'est-à-dire indépendamment des usages postérieurs qui en sont faits], elle a permis l'évaluation et l'optimisation des systèmes de RAP pendant de longues années » (Ben Jannet, 2015). Le WER permet en particulier de comparer différents systèmes entre eux (notre besoin) ainsi que d'évaluer leur amélioration dans le temps. Les scores compilés sur plusieurs années, même si les procédures mises en place sont forcément disparates, permettent globalement de voir les progrès importants réalisés par les outils de transcription. En utilisant le WER, nous pouvons également comparer nos résultats à ceux obtenus par d'autres.

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.

Normalisation des fichiers

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.

Implémentation sous R

La normalisation ainsi que le calcul des WER ont été réalisés avec le logiciel R (R Core Team, 2020). Nous avons utilisé une fonction développée par Jens Wäckerle pour le calcul des WER. La fonction fait partie du package wersim disponible sur Github[25]. Une petite mise à jour a été nécessaire pour pouvoir l'utiliser. Le script R est disponibleAnnexe 2 — Script R.

Résultats

Le Tableau V-1 présente les résultats pour chaque fichier et chaque plateforme, en s'inspirant du modèle de présentation réalisé par Tim 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. Chaque cellule est coloriée en fonction de la proximité du score obtenu avec la médiane (rouge : bien moins bon que la médiane, vert : bien meilleur, jaune : proche de la médiane). Plus le score WER est proche de 0, mieux c'est.

	Manuelle	Médiane	Happy Scribe	Go Transcribe	Sonix	Vocapia	Video Indexer	YouTube	Headliner	Vocalmatic
Physionomie (texte lu)	0,6	19,5	14,8	14,8	16,6	18,1	20,8	28,4	28,2	28,5
Comptines (monologue)	1,6	14,8	11,9	12,4	11,1	14,2	18,2	15,4	31,1	26,6
Camille (entretien face à face)	18,9	50,0	51,8	54,2	48,3	28,1	34,0	36,4	76,1	75,5
Harmonie (discussion spontanée)	12,7	88,4	86,2	86,2	90,5	57,6	79,8	107,0	222,6	244,6

Tableau V-1 comparaison des plateformes selon 4 fichiers, par calcul du WER

[25] https://rdrr.io/github/jenswaeckerle/wersim/src/R/wer.R

Une première observation est que quel que soit le type d'enregistrement et quelle que soit la plateforme, 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 humain.

Néanmoins, trois grands groupes se distinguent :

1. des plateformes particulièrement performantes pour les discours simples (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.

Comparaison avec d'autres études

Dans son article de blog mis à jour en 2018, Tim Bunce (2018) observe sur son corpus en anglais que « les meilleurs services de transcription automatique obtiennent tous un score de 10 à 16, la plupart se situant autour de 12. Les scores du niveau suivant sont à peu près deux fois plus élevés : 22-28. Il semble probable que les systèmes de premier niveau utilisent des technologies plus modernes » (traduction par nos soins).

Les résultats que nous obtenons avec les deux corpus a priori les plus adaptés à un traitement automatique (`Physionomie` et `Comptines`) correspondent bien aux valeurs obtenues par Bunce, entre 12 et 16.

Nous observons par contre une importance disparité sur les scores obtenus pour nos transcriptions manuelles qui sont soit bien meilleures sur les monologues, soient moins bonnes lorsqu'il s'agit des corpus présentant des difficultés légères (`Camille`) ou fortes (`Harmonie`). Notons que l'étude de Bunce à laquelle nous faisons référence ne repose que sur des monologues (avec des accents et dictions variés) sauf le fichier F18 avec un homme et deux femmes qui parfois se coupent la parole. Dans ce cas, la transcription manuelle passe à un WER de 11 ce qui correspond davantage à nos résultats. Notons également que nous ne tenons pas compte non plus des différences de performances qui pourraient exister entre la reconnaissance automatique de différentes langues (anglais VS français).

Quant aux valeurs les plus « mauvaises » qu'il ait obtenu, il évoque des WER à 30-40 ce qui correspond là aussi assez bien aux valeurs obtenues (si on élimine `Harmonie`). Dans ce cas, Headliner et Vocalmatic sont plus mauvaises que les chiffres qu'il annonce.

Ainsi, si le WER permet théoriquement de comparer des études entre elles, les limites liées aux écarts méthodologiques et à l'utilisation de diverses langues rendent hasardeux tout véritable approfondissement de ces comparaisons. Il nous a toutefois semblé intéressant de pouvoir – même succinctement – confronter nos résultats à d'autres et y trouver une cohérence globale.

Mais comment aller au-delà du calcul du WER ? Les résultats obtenus ne nous disent rien sur la nature des erreurs générées. Nuisent-elles à la compréhension ? Sont-elles plus ou moins longues à corriger ?

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

Introduction

Le WER calculé précédemment nous a permis de nous faire une idée globale de la correspondance qu'il existe entre nos transcriptions de référence (REF) et les transcriptions produites par les plateformes (RES). Nous avons souhaité compléter cet indicateur par un parcours systématique des transcriptions produites et des erreurs qui s'y trouvent.

Cette étape ne vise pas à établir un nouvel indicateur à proprement parler. Si elle peut y participer, elle cherche davantage à porter un regard critique sur les objets textuels issus des transcriptions produites. Dans la lignée des réflexions menées sur l'instrumentation de la linguistique (Habert, 2005), elle invite le lecteur à s'interroger sur les biais introduits par la transcription automatique et sur l'objet langagier ici manipulé. Quelles difficultés représente encore aujourd'hui l'ambiguïté phonétique à laquelle sont confrontés les instruments ? Comment font-ils face à la diversité des situations de communication, aux chevauchements de parole, aux bruits parasites et aux voix atypiques ? Quelles sont les conséquences de cette variation sur la nature des objets textuels qu'ils produisent ? Peut-on envisager de les intégrer à une chaîne de traitements plus complexes dans laquelle des post-traitements amélioreraient à moindres frais la qualité des transcriptions ?

Loin de prétendre répondre à ces questions, le travail réalisé ici propose 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 RES produits pour les trois sous-corpus `Physionomie`, `Comptines` et `Camille`[26].

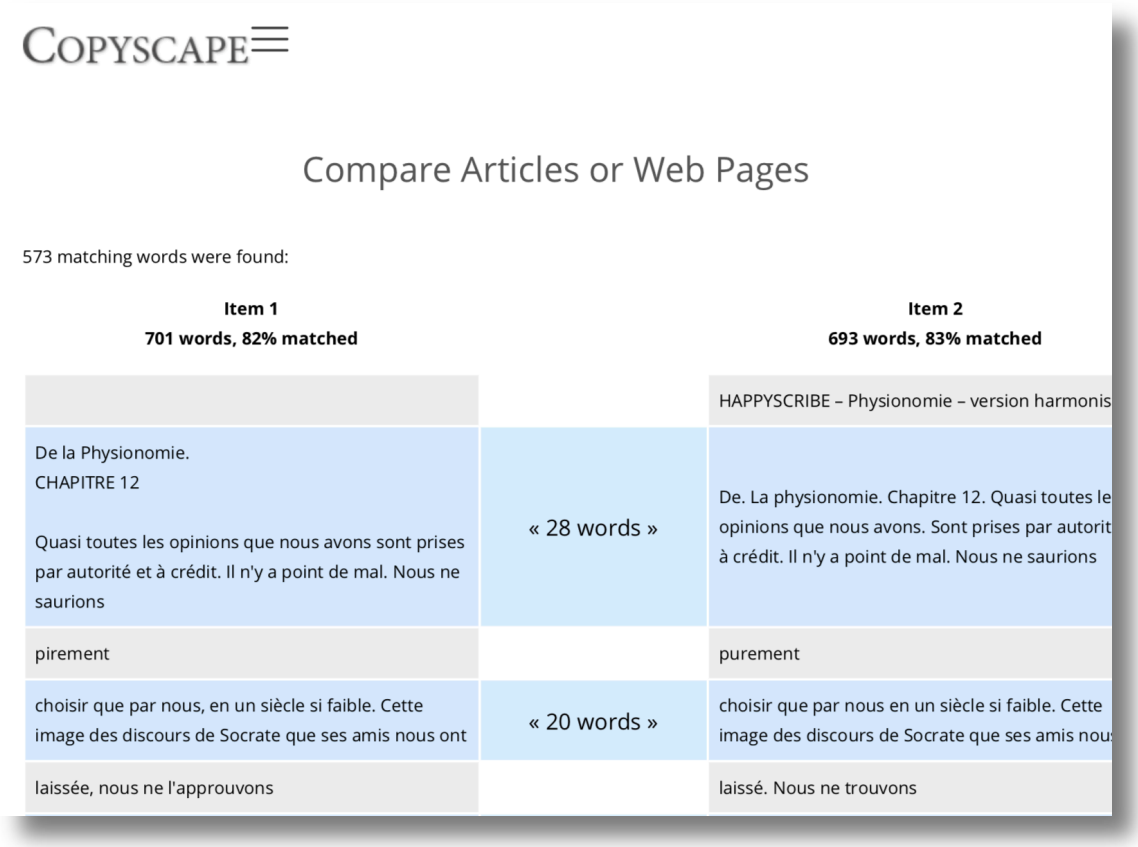
Nous expliciterons ici la méthode mise en œuvre pour élaborer notre typologie, avant de décrire les différentes catégories définies. Enfin, nous partagerons avec vous une caractérisation des fichiers RES analysés à travers cette grille de lecture.

Apport des outils antiplagiat

Afin d'amorcer l'étude des erreurs produites par les plateformes, plusieurs solutions techniques ont été envisagées. Nous souhaitions utiliser une solution simple à prendre en main, gratuite, accessible à chacun et chacune d'entre nous et dont les résultats pouvaient facilement être parcourus et interprétés. Plusieurs solutions ont été proposées et testées. C'est finalement Copyscape, un outil de détection de plagiat, qui a été retenu. Copyscape propose un outil de comparaisons de textes en ligne[27]. La version gratuite est limitée en matière de fonctionnalités et de nombre de comparaisons offertes par jour. Une version payante plus complète est également disponible. Le service gratuit se présente sous la forme d'une page web à partir de laquelle il est possible de comparer deux sources (pages web ou fragments de texte). Bien que l'éditeur n'indique pas explicitement sur quels algorithmes repose le service,

[26] Le sous-corpus `Harmonie` a été exclu au cours de la procédure en raison du temps de traitement induit par sa complexité.
[27] À cette adresse : https://www.copyscape.com/compare.php

une simple observation des résultats nous donne des indications sur son fonctionnement. En résumé, l'outil effectue une comparaison mot à mot des deux sources qui lui sont soumises. Il 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. Les signes de ponctuation, la casse et les sauts de ligne ne sont pas pris en compte. Ainsi la chaîne *Il passa le Week-End chezz lui* est considérée comme identique à *il passa, le week end chezz lui*. À la fin d'un segment commun, Copyscape reprend son parcours en alternant lignes de mots identiques et lignes de mots différents. Au final, Copyscape retourne un rapport sous forme de tableau composé de 3 colonnes, comme le montre la Figure VI-1. L'en-tête de ce rapport comporte quelques statistiques, le nombre de mots contenus dans les segments communs et pour chaque source le nombre total de mots ainsi que le ratio mots communs/nombre de mots.



COPYSCAPE

Compare Articles or Web Pages

573 matching words were found:

Item 1 701 words, 82% matched		Item 2 693 words, 83% matched
		HAPPYSCRIBE – Physionomie – version harmonis
De la Physionomie. CHAPITRE 12 Quasi toutes les opinions que nous avons sont prises par autorité et à crédit. Il n'y a point de mal. Nous ne saurions	« 28 words »	De. La physionomie. Chapitre 12. Quasi toutes le opinions que nous avons. Sont prises par autorit à crédit. Il n'y a point de mal. Nous ne saurions
pirement		purement
choisir que par nous, en un siècle si faible. Cette image des discours de Socrate que ses amis nous ont	« 20 words »	choisir que par nous en un siècle si faible. Cette image des discours de Socrate que ses amis nou
laissée, nous ne l'approuvons		laissé. Nous ne trouvons

Figure VI-1 Exemple de sortie de Copyscape

Une fois l'ensemble des fichiers harmonisés, selon la procédure décrite dans la section III, chacun d'entre nous a exécuté Copyscape pour comparer les fichiers RES qu'il venait d'harmoniser, aux fichiers REF correspondants. Le résultat de chacune de ces comparaisons a été enregistré sous deux formats :

- un fichier au format PDF permettant de conserver une trace de la sortie exacte de l'outil ;
- un classeur Excel, dans lequel le tableau généré par Copyscape a été copié-collé, pour mener le travail d'analyse des erreurs.

Cette étape a été réalisée sur l'ensemble de notre corpus.

Résultats bruts Copyscape

Le Tableau VI-1 montre les résultats bruts de Copyscape. Plus précisément, il présente la proportion de mots de chaque transcription RES (représentant chacune un couple plateforme/sous-corpus) considérée comme un plagiat de la transcription REF correspondante.

La mise en couleur permet de visualiser rapidement les plateformes qui offrent les meilleures performances pour un sous-corpus donné (en vert foncé) et celles qui offrent les plus mauvaises (en rouge). Un dégradé du vert clair à l'orangé, en passant par le jaune, permet de visualiser la répartition des résultats intermédiaires.

	Médiane	Happy Scribe	Go Transcribe	Sonix	Vocapia	Video Indexer	YouTube	Headliner	Vocalmatic
Physionomie	78 %	83 %	83 %	81 %	79 %	77 %	72 %	68 %	66 %
Comptines	88 %	90 %	89 %	90 %	88 %	82 %	88 %	79 %	80 %
Camille	74 %	79 %	77 %	77 %	77 %	65 %	71 %	65 %	67 %
Harmonie	63 %	64 %	65 %	61 %	59 %	54 %	56 %	67 %	67 %

Tableau VI-1 Proportion mots RES communs à mots REF selon Copyscape

La lecture de ce tableau nous amène rapidement à formuler deux constats :

- l'ensemble des plateformes fournissent leur meilleure transcription pour le sous-corpus `Comptines`, un monologue enregistré dans de bonnes conditions ;
- les performances varient fortement d'un sous-corpus à l'autre.

Nous observons également que les transcriptions d'Happy Scribe sont les plus proches des REF pour trois sous-corpus sur quatre. Sans grande surprise, le sous-corpus `Harmonie`, choisi parce qu'il comporte de nombreux locuteurs, des bruits ambiants et des chevauchements de paroles, semble poser plus de difficultés à la majeure partie des plateformes. Seuls Headliner et Vocalmatic retournent, pour ce sous-corpus, des transcriptions avec une proportion de mots communs similaire à celles qu'ils retournent pour `Physionomie` et `Camille`.

Attention, cependant, la proportion est ici calculée sur la taille des transcriptions produites par les plateformes (texte RES). Elle nous informe de la part de mots des RES qui sont identiques à ceux présents dans les REF pour les passages effectivement transcrits, mais non de leur nombre. Si l'on s'intéresse au Tableau VI-2, qui présente la taille de chacune des transcriptions en jeu, on constate que la taille des transcriptions varie davantage pour les sous-corpus `Camille` et `Harmonie` que pour les deux autres. Dans le cas précis d'`Harmonie`, les transcriptions proposées par Headliner et Vocalmatic comportent un nombre de mots plus de deux fois inférieur à celui de la transcription de référence. À l'inverse, la transcription proposée par Vocapia est la plus proche, en nombre de mots, de la version de référence. Cette disparité invite à une analyse plus fine des résultats couplée à une lecture qualitative des transcriptions : les transcriptions les plus couvrantes comportent-elles un taux d'erreurs légèrement supérieur ? Les plus partielles sont-elles les plus fiables ? Ces premiers résultats ne nous permettent pas de le déterminer. Si une telle corrélation se trouvait observée par la suite, il reviendrait au lecteur ou à la lectrice de préférer l'une ou l'autre des solutions,

en fonction de l'objet de ses recherches et des projets pour lesquels il ou elle réalise des transcriptions.

	REF	Médiane	Happy Scribe	Go Transcribe	Sonix	Vocapia	Video Indexer	YouTube	Headliner	Vocalmatic
Physionomie	701	694	693	693	692	700	694	658	702	701
Comptines	837	805	805	810	805	807	799	807	736	754
Camille	661	499	490	487	508	619	643	609	443	432
Harmonie	794	483	491	487	478	634	527	445	255	237

Tableau VI-2 Taille des transcriptions en nombre de mots

Copyscape nous fournit donc deux indicateurs de la qualité lexicale globale 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.

Il nous fournit également un format intéressant pour servir de base à l'analyse plus fine que nous souhaitions mener et que nous présenterons dans les sections suivantes.

Grille d'analyse des erreurs de transcription

Les résultats chiffrés fournis par Copyscape nous donnent une première idée de la fidélité lexicale des transcriptions obtenues à l'aide des différentes plateformes. Tout comme le WER, ils ne nous disent cependant rien sur les erreurs introduites par la transcription automatique et très peu sur la propension des plateformes à ne pas transcrire certains passages ou à ajouter des mots.

La sortie fournie par l'outil nous permet cependant un gain de temps important en identifiant les passages sur lesquels nous concentrer. Ce sont ces passages, considérés comme ne relevant pas du plagiat, que nous avons choisi d'annoter manuellement pour établir notre typologie.

> **Remarque** : à partir de maintenant, nous appellerons **mot**[28] 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 1 mot).
> **Problème** : Comment compter 9 *h 30* ? Un mot ou trois ? Nous en compterons ici trois, car nous avons privilégié la convention de notation 9 *heures 30* dans les transcriptions de référence.

[28] 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.

Organisation de l'annotation

À partir de cette étape, le travail n'a plus été réalisé par l'ensemble des membres du groupe de travail, mais il a été confié à deux personnes seulement, pour s'assurer d'un jugement le plus continu possible. Le travail s'est organisé de la façon suivante :

- Une première personne a parcouru l'ensemble des huit fichiers comparant le fichier REF d'un premier sous-corpus (`Physionomie`) aux fichiers RES correspondants. Elle a opéré une classification des mots de fichiers RES en six catégories (mots communs, mauvais lexique, mots absents, mots ajoutés, erreurs de flexion, autres).
- Elle a rédigé une version 0 d'un guide d'annotation, décrivant les différentes catégories.
- Une seconde personne, après lecture de la version 0 du guide, a relu les classifications produites par la première pour avis et échanges.
- Les deux personnes se sont ensuite réparti les 3 fichiers REF restants : `Camille`, `Comptines` et `Harmonie`, elles ont opéré une première classification, puis se sont échangé les fichiers pour relecture et discussion.
- Le guide d'annotation a évolué au fur et à mesure de l'analyse des fichiers, pour aboutir à une classification en neuf catégories (mots de RES identiques à REF, substitution avec proximité phonétique, substitution sans proximité phonétique, mots de REF absents de RES, mots de RES qui ne correspondent à rien dans REF [ajout], nombre de mots REF dans les passages sans aucun lien, nombre de mots RES dans les passages sans aucun lien, erreurs de flexion, autres).

Typologie des erreurs de transcription

Rappel : la mesure WER est basée sur trois opérations : la substitution (`sub.`), l'ajout (`aj.`) et la suppression (`sup.`). Pour un énoncé transcrit *où est sa plus juste et laborieuse besogne* dans un fichier REF et *où est l'état ça pue juste et plus laborieux spécial* dans un fichier RES, les opérations se répartissent ainsi :

où	***est***		***sa***	***plus***	***juste***	***est***		***laborieuse***	***besogne***
où	*est*	*l'état*	*ça*	*pue*	*juste*	*et*	*plus*	*laborieux*	*spécial*
		aj.	*sub.*	*sub.*		*sub.*	*aj.*	*sub.*	*sub.*

Comme nous l'avons évoqué précédemment, le travail manuel réalisé ici a pour vocation de nous permettre 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 la typologie des erreurs s'est affinée au fur et à mesure de nos relectures. Elle a fait l'objet de la rédaction d'un guide d'annotation, disponible en annexe (Annexe 3 — Guide d'annotation). Ce guide, associé à des relectures croisées, nous a permis d'uniformiser la manière dont nous utilisions chaque catégorie. Ce travail a été réalisé à distance, par deux annotateurs n'ayant jamais travaillé ensemble au préalable. Ces conditions particulières ont nécessité de nombreux échanges et ajustement, à la fois par mail et par visioconférence. Bien qu'ayant été mené avec rigueur, il n'a pas prétention à permettre une évaluation quantitative directe des différentes transcriptions. Il offre un

regard complémentaire à ceux des autres méthodes de comparaison des plateformes proposées dans ce document et invite le lecteur à s'approprier ce regard.

Il est important de comprendre que la classification menée ici nous a naturellement amenés à dénombrer davantage de mots identiques que ceux présentés dans la section **Résultats bruts Copyscape**. Si l'on s'intéresse à l'extrait présenté dans la Figure VI-2, on voit qu'il ne comporte pas trois mots consécutifs identiques. Il n'est donc pas comptabilisé par Copyscape comme une zone de plagiat potentielle. Cependant, on voit bien que la transcription proposée comporte un mot identique à ceux de la transcription de référence. Nous avons annoté de tels mots comme étant des **mots de RES identiques à REF** et ils sont venus grossir le décompte fourni par Copyscape.

701 words, 68% matched		702 words, 68% matched	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	mots de REF absents de RES	mots de RES qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	erreurs de flexion	autre
préceptes, qui		précepte qui est	1	0		0	1			1	0

Figure VI-2 Transcription Headliner — « préceptes, qui »

De plus, l'annotation manuelle nous permet d'aborder la variation avec souplesse. Ainsi, *minutes* et son abréviation *min.* seront jugés comme étant des transcriptions identiques, tout autant que *7* et *sept* ou *essaye* et *essaie*.

L'exemple de la Figure VI-2 comporte une occurrence de *précepte* intéressante. Comme on peut le voir, ce mot ne diffère de celui présent dans la transcription de référence que par son *s* final, marque du pluriel pour ce nom commun. Nous avons souhaité créer une catégorie particulière pour ce type d'erreurs, que nous désignons sous le terme d'**erreurs de flexion**. Si de telles erreurs peuvent induire des différences de sens importantes (*celles qui coulent échappent* vs *celle qui coule échappe*), nous considérons qu'elles nuisent moins à la compréhension générale du discours que lorsque l'unité lexicale présente dans la transcription n'est pas la bonne. Lors de l'analyse des transcriptions, nous chercherons à savoir s'il est possible de réduire leur nombre à l'aide d'un post-traitement.

En créant cette catégorie, nous introduisons une première nuance à l'intérieur de ce que la mesure WER considère comme des substitutions.

L'**erreur de flexion** est toujours strictement équivalente en nombre de mots à ce que décompte le WER. Elle se distingue en cela des deux sous-types de substitution que nous avons introduits : la **substitution avec proximité phonétique** et la **substitution sans proximité phonétique**.

Si l'on s'intéresse au couple de transcriptions *il ne monta rien, mais ravala/Ils ne m'ont pas rien, mais Ravana*, visible dans la Figure VI-3, on voit que la transcription proposée par la plateforme comporte deux erreurs. Là où la transcription de référence indique *monta* (un mot) la plateforme propose *m'ont pas* (deux mots), là où le fichier REF contient *ravala*, le fichier RES contient *Ravana*, un mot partout. Ces erreurs nous semblent justifiées par une forte proximité phonétique. Elles sont de celles que nous désignons comme étant des cas de **substitution avec proximité phonétique**.

Item 1		Item 2	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	mots de REF absents de RES	mots de RES qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	erreurs de flexion	autre
il ne monta rien, mais ravala		Ils ne m'ont pas rien, mais Ravana	3	3		0	0			1	0

Figure VI-3 Transcription Sonix — « il ne monta rien, mais ravala »

Notre point de vue étant axé sur les objets textuels produits par les instruments de transcription automatique et par une comparaison manuelle des fichiers **REF** et **RES** notre méthode de comptage n'est, à partir d'ici, plus comparable directement à celle de la mesure WER. En effet, là où la mesure WER comptabiliserai deux substitutions et un ajout :

il	***ne***	***monta***		***rien,***	***mais***	***ravala***
il	*ne*	*m'ont*	*pas*	*rien,*	*mais*	*Ravana*
		sub.	*aj.*			*sub.*

Nous considérons que nous sommes en présence de trois mots en jeu dans des substitutions :

il	***ne***	***monta***		***rien,***	***mais***	***ravala***
il	*ne*	*m'ont*	*pas*	*rien,*	*mais*	*Ravana*
		sub.	*sub.*			*sub.*

En l'absence complète de proximité phonétique entre les deux transcriptions proposées (REF et RES), nous avons choisi de distinguer deux situations : les erreurs de **substitution sans proximité phonétique** et les **passages sans aucun lien**.

À la lecture des sous-corpus `Physionomie` et `Comptines`, lorsque nous rencontrions des cas d'unités lexicales sans lien évident entre REF et RES, nous parvenions toujours à trouver une correspondance entre le texte de REF et celui de RES. Ainsi, dans le cas du couple de transcription *Nous n'apercevons/nous fait recevant* de la Figure VI-4, nous pouvons concevoir que ce que l'instrument a transcrit *fait recevant* correspond au segment transcrit *n'apercevons* par le transcripteur humain. Si nous percevons bien une proximité phonétique entre *recevant* et la partie postérieure d'*apercevons*, nous avons plus de mal à admettre que la substitution de *n'* et de la partie antérieure de *apercevons* devienne ici *fait*. Nous lui appliquons donc la catégorie **substitution sans proximité phonétique**.

699 words, 66% matched		701 words, 66% matched	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	mots de REF absents de RES	qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	erreurs de flexion	autre
Nous n'apercevons		nous fait recevant	1	1	1	0	0			0	0

Figure VI-4 Transcription Vocalmatic — « nous n'apercevons »

De nouveaux cas de figure se sont présentés à nous dans les transcriptions proposées pour le sous-corpus `Camille`. Que faire face à un couple de transcriptions tel que *Alors, il y a une époque où j'allais à Mérignac, maintenant j'y vais moins/Johnny et David*

parce que présenté dans la Figure VI-5 ? Peut-on encore considérer ici qu'il y a la moindre substitution ? Faut-il comptabiliser autant de mots absents que de mots contenus dans RES et de mots ajoutés que de mots contenus dans REF ? Nous avons préféré proposer ici un couple de catégories à part. Ce couple de catégories **nombre mots REF**, **nombre mots RES**, associé à la mention **passages sans aucun lien**, nous permet de comptabiliser de telles erreurs de transcription comme un fait spécifique.

661 words, 44% matched		443 words, 65% matched	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	mots de REF absents de RES	qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	erreurs de flexion	autre
Alors, il y a une époque où j'allais à Mérignac, maintenant j'y vais moins		Johnny et David parce que	0	0		0	0	14	5	0	0

Figure VI-5 Transcription Headliner — Passage sans lien

Nous distinguons ainsi ce cas de figure de l'absence totale de transcription, présentée dans la Figure VI-6. Dans le couple *où j'allais à Mérignac, maintenant j'y/ou j'allais, maintenant je*, aucun mot de RES ne correspond à *à Mérignac*. Nous proposons donc de catégoriser ces deux mots comme **mots de REF absents de RES**. Cette distinction peut permettre, lors d'un test préalable au choix d'un instrument pour un projet donné, d'évaluer la stratégie de la plateforme face à un corpus d'une certaine qualité : produit-elle la transcription la plus couvrante possible ou une transcription plus partielle ? Si la transcription est très couvrante, comporte-t-elle de nombreux passages sans aucun lien ou majoritairement des substitutions raisonnablement alignables ? Sera-t-il possible de l'améliorer significativement à l'aide de post-traitement ou faudra-t-il exclusivement compter sur une correction manuelle ?

661 words, 72% matched		619 words, 77% matched	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	mots de REF absents de RES	mots de RES qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	erreurs de flexion	autre
où j'allais à Mérignac, maintenant j'y		ou j'allais, maintenant je	2	1		2	0			0	1

Figure VI-6 Transcription Vocapia — « où j'allais à Mérignac, maintenant j'y »

Parallèlement, cette distinction peut permettre d'évaluer la propension des plateformes à proposer des mots là où le transcripteur humain a considéré qu'il n'y en avait aucun, comme c'est le cas dans le couple *me raconterais/peux me raconter* présenté dans la Figure VI-7[29].

Item 1 661 words, 63% matched		Item 2 643 words, 65% matched	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	erreurs de flexion	mots de REF absents de RES	mots de RES qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	autre
me raconterais		peux me raconter	1	0	0	0	1	0	0	1	0

Figure VI-7 Transcription Video Indexer — « me raconterais »

Pour finir, nous avons créé une catégorie **autre**, dans laquelle nous avons regroupé différents cas de figure que nous ne souhaitions pas répertorier ailleurs. Ces cas se sont avérés peu nombreux. Ils correspondent majoritairement à des couples de transcriptions tels que *j'y/je* dans la Figure VI-6, pour lesquels la segmentation en mots que nous avons choisie ne nous permettait pas de classer de manière satisfaisante les

[29] Il n'échappera pas au lecteur que les deux transcriptions de ce couple sont sémantiquement équivalentes. Nous n'avons pas pris la peine de créer une catégorie spécifique pour de tel cas, mais cela pourrait s'avérer pertinent en fonction des objectifs du projet pour lequel le test est réalisé.

mots en présence. Dans ce cas précis, nous aurions aimé pouvoir dire que *j'* et *je* correspondent à la même unité lexicale, tandis que *y* est absent de REF.

Cas particulier de Harmonie : des passages inaudibles ou incertains

Cette première phase de l'analyse a été réalisée avec une contrainte temporelle forte. Dans ce laps de temps, l'analyse complète du sous-corpus `Harmonie`, qui cumule plusieurs difficultés (locuteurs multiples, chevauchements de parole, bruit ambiant) s'est avérée impossible.

`Harmonie` n'est cependant pas un cas exceptionnel que nous pourrions simplement exclure de notre étude. Dans nos recherches respectives, il nous arrive d'avoir à transcrire des corpus qui, tout comme ce sous-corpus, contiennent des passages difficiles. Il en va ainsi de ceux qui contiennent des chevauchements multiples (entame d'un nouveau chevauchement au sein d'un chevauchement en cours) ou des discussions en marge de la conversation principale. C'est également le cas pour les enregistrements qui comportent des bruits de fond importants provoqués par un événement externe à la conversation (bruits de travaux, circulation automobile, bruit de ventilation intérieure, tapotements…) ou dont la qualité de la captation est mauvaise (problème de micros, avec les enregistreurs…).

Face à de tels passages, les plateformes poursuivent leur travail de reconnaissance automatique en transcrivant les mots qu'elles « perçoivent » au fil des énoncés. Ce qui conduit à une transcription atomisée et largement tronquée.

Dans le cas de notre contexte méthodologique, cette atomisation rend la comparaison mot à mot difficile du fait du décalage qu'elle induit entre REF et RES. Premièrement, ce décalage se traduit par une différence importante de nombre de mots entre les transcriptions, comme le montre la Figure VI-8.

REF	RES
oui enfin ils sont reconfirmés ils réprés- ils sont reconfirmés dans le tard reconfirmés ou modifiés ils sont reconfirmés enfin voilà est-ce qu'il un	représente, reconfirmé, reconfirmé.

Figure VI-8 Transcription Sonix — « segments incertains »

Deuxièmement, si la sortie produite par Copyscape fournit un bon alignement des segments identiques, même après des passages de ce type, la mise en correspondance ligne à ligne des passages difficiles est rarement satisfaisante, comme on peut le voir dans la Figure VI-9.

REF	RES
	Changement de président de résidence du présdent Moneim.
oui d'accord qu'il eut changé ou pas ben on commence par le donc on commence par président président vice-président et trésorier donc président donc moi-même	

Figure VI-9 Transcription Sonix — « Copyscape, décalage alignement REF/RES »

Se pose ensuite la question de l'annotation de ces passages. Nous proposons ici une piste qui permet de coller à la typologie des erreurs présentée dans la section précédente, tout en signalant la spécificité à laquelle on se trouve confrontés. Faute de temps, cette piste n'a cependant pas été mise en œuvre.

Les passages qui comportent des chevauchements multiples, des discussions parallèles, des bruits de fond importants ou qui s'avèrent être de mauvaise qualité sonore pourraient être annotés **segments ambigus**. Nous jugeons ces passages trop complexes pour être transcrits sans équivoque. Cela ne veut pas forcément dire qu'ils sont inaudibles, mais simplement qu'il devient difficile, même pour un transcripteur humain expérimenté, de le transcrire sans risque d'erreur.

Pour ces segments, il nous semblerait pertinent d'ajouter des informations temporelles, sous la forme de balises temporelles de début et de fin et d'une indication de durée, exprimée en secondes, comme le montre la Figure VI-10. Ces informations nous permettront d'estimer l'ampleur du problème sur le sous-corpus `Harmonie`.

(2 h 51 - 5 s - segment ambigu : passage avec chevauchements et segments incertains)

REF	RES
enfin moi ils sont reconfirmés ils réprése — ils sont reconfirmés dans le tard reconfirmés ils sont reconfirmés ou modifiés enfin voilà	représente, reconfirmé, reconfirmé.

(2 h 56 – Fin)

Figure VI-10 Transcription Sonix — Notation des segments ambigus

Il serait ensuite possible d'appliquer à ces passages un traitement similaire à celui spécifié pour les passages qu'il semble difficile d'aligner dans n'importe quel corpus :

- Compter le nombre de mots de RES identiques aux mots de REF
- Relever le nombre de mots de REF contenus dans ces passages
- Relever le nombre de mots de RES contenus dans ces passages

Cependant, étant donné la difficulté qu'ils représentent, même pour un transcripteur humain expérimenté, nous envisagerions d'exclure ces passages des décomptes principaux.

Analyse des résultats

Une fois notre typologie d'erreurs mise en place, nous avons souhaité regarder notre corpus à travers cette grille d'analyse. Nous pourrions ainsi voir quelles tendances elle permet de dégager et s'il est possible d'en faire un outil de comparaison des sorties de différentes plateformes.

L'analyse que nous proposons n'est cependant pas à prendre pour argent comptant dans le but de privilégier ou d'exclure telle ou telle plateforme. Les données que nous avons analysées sont en très petites quantités. Il s'agit de textes comportant de 436 à 818 mots seulement. Nous invitons le lecteur à prendre cette analyse comme la présentation de pistes pour lui permettre de réaliser ses propres tests sur un extrait du corpus qu'il souhaite transcrire automatiquement.

Rappelons-le, le corpus à notre disposition est constitué de quatre sous-corpus, dont trois seulement ont été entièrement traités. Chacun de ces trois sous-corpus a été transcrit automatiquement par huit plateformes. Nous disposons donc de trois sous-ensembles chacun composé d'un fichier REF et de 8 fichiers RES, soit un total de 24 fichiers RES analysés.

Ces sous-corpus appartiennent à des genres discursifs et interactionnels distincts, dont nous rappelons ici les principales caractéristiques :

`Physionomie` est un texte littéraire issu des essais de Montaigne (Livre III/Chapitre 12). Dans le cadre de cette étude, nous avons exploité la version en français modernisée. Le texte est lu par un narrateur professionnel. La qualité de l'enregistrement est excellente.

`Comptines` est extrait de l'enregistrement d'un cours magistral universitaire. On peut donc le classer dans la catégorie des monologues académique. Le registre de langue de l'intervenante peut être qualifié de soutenu. Sa diction est bonne et peu d'onomatopées ou d'hésitations ponctuent le discours. La qualité du signal est correcte, sans bruits de fond prononcés.

`Camille` est extrait d'un entretien de recherche entre deux locuteurs, dont l'un rencontre des difficultés d'énonciation dues à un handicap physique. Les conditions d'enregistrement sont relativement bonnes, sans bruit parasite.

Physionomie

L'ensemble des plateformes que nous avons testées retournent des performances honorables pour le sous-corpus `Physionomie`. Comme le montre la Figure VI-11, la proportion de mots identiques à ceux du fichier REF dans chacun des fichiers RES est relativement élevée (située entre 75,07 et 88,52 %). Nous attribuons ces honnêtes performances aux caractéristiques générales de l'élocution du locuteur : un professionnel qui maîtrise la plupart des aspects prosodiques du discours (rythme, intensité, intonation…). De plus, la situation particulière de texte lu exclut tout phénomène propre à une conversation spontanée comme les hésitations, les troncations ou encore les chevauchements de paroles. Enfin, le signal est lui-même

d'excellente qualité et ne comporte aucun phénomène parasite (coupures, bruits de fond…).

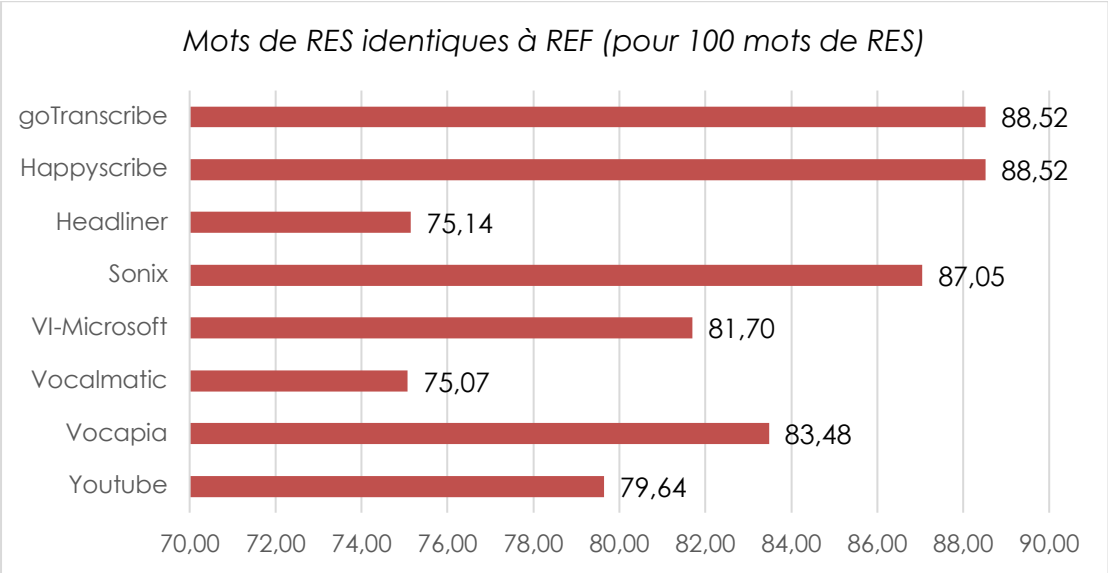


Figure VI-11 Proportion de mots identiques à REF dans chaque fichier RES — Physionomie

On constate également, à la lecture de la Figure VI-12, que peu de mots présents dans la transcription de référence ont été considérés comme n'ayant pas été transcrits par les plateformes. À part YouTube, pour qui la proportion ne dépasse pas les 92,3 %, les instruments ont proposé une transcription pour au moins 98 % des mots. Pour autant, YouTube n'est pas la plateforme qui offre la plus grande proportion de mots identiques entre fichiers RES et REF.

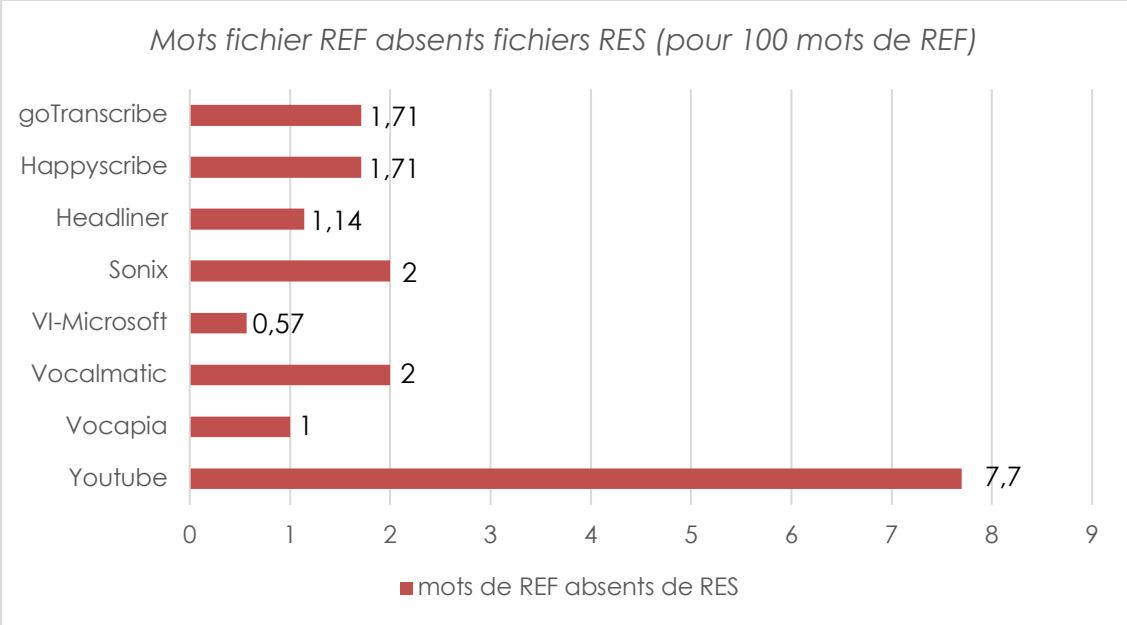


Figure VI-12 Proportion de mots de REF considérés comme n'ayant pas été transcrits — Physionomie

Revenons-en aux mots des fichiers RES identiques aux mots du fichier REF. Si l'on ajoute à ces mots identiques les mots identifiés comme comportant une erreur de flexion (Figure VI-13), nous atteignons des taux de reconnaissance qui se situent entre 82,91 et 94,04 %. Les erreurs de flexion sont ici relativement nombreuses. Elles représentent entre 5,52 % et 9,72 % des mots transcrits et leur correction permettrait une amélioration moyenne des performances de 7,40 %. Elle ne changerait cependant rien à la répartition des performances par plateformes pour ce sous-corpus.

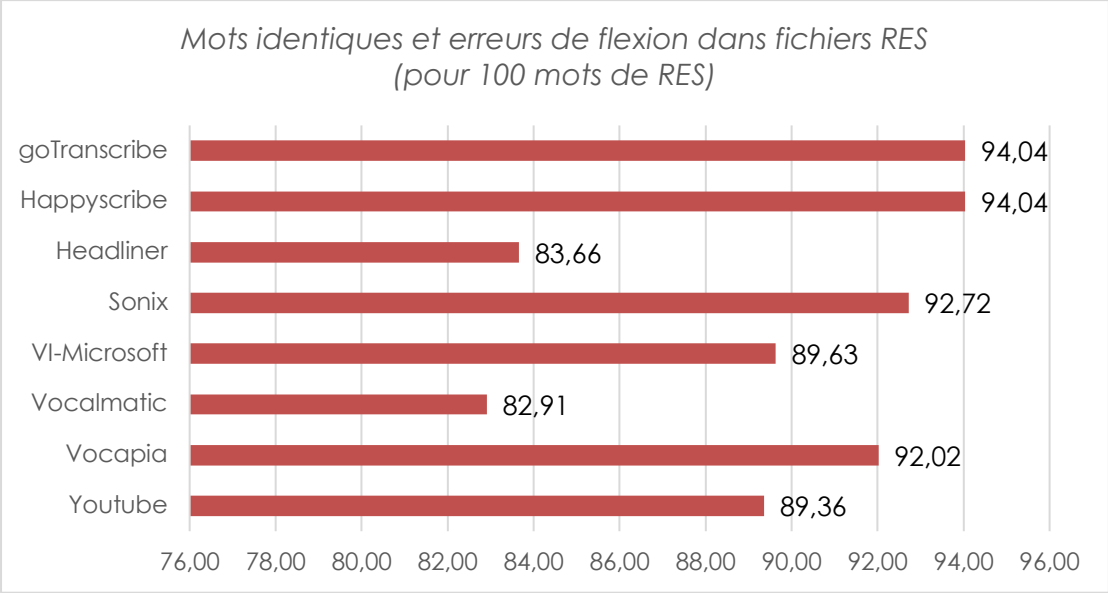


Figure VI-13 Proportion de mots identiques à REF ou ne variant que par la flexion dans chaque fichier RES — Physionomie

Pour aller un peu plus loin dans cette observation, nous pouvons nous demander s'il est possible d'envisager un post-traitement automatique qui permettrait cette correction. Nous avons constaté que les erreurs de flexion rencontrées concernent principalement des adjectifs, des verbes conjugués et des participes passés. Examinons-en quelques-unes :

REF	→	**RES**
la naïveté et la simplicité *échappent*	→	la naïveté et la simplicité *échappe*
nous qui estimons *plates et basses* toutes celles	→	nous qui estimons *plate et basse* toutes celles
c'est une allure *tendue*	→	c'est une allure *tendu*
d'un pas *mol*	→	d'un pas *molle*
aux plus *épineuses* traverses qui se *puissent*	→	*au* plus *épineuse traverse* qui se *puisse*

Certaines relations syntaxiques sont ici très simples à identifier, comme celle qui relie l'adjectif *tendue* au nom *allure* qui le précède. D'autres, telles que celles mettant en jeu une coordination, sont moins évidentes à détecter. Pour autant tous les cas d'erreurs présentés peuvent être corrigés à partir du texte produit par les plateformes, sans avoir recours à la version audio du sous-corpus.

Il est intéressant de noter que l'antéposition de certains adjectifs épithètes semble poser problème aux plateformes. Ainsi, là où la transcription de référence propose *de plus vulgaires et connues actions des hommes*, aucune des huit plateformes ne parvient à accorder *connues* à *actions*[30]. Nous nous appuyons ici sur les travaux de Benzitoun (2014) et la fiche de synthèse de Forsgren (2016), pour émettre l'hypothèse qu'il s'agit là d'une difficulté relevant du genre particulier du sous-corpus `Physionomie`. La position des adjectifs épithètes en français étant davantage contrainte à l'oral qu'à l'écrit, nous pouvons supposer que les plateformes de transcription n'ont pas rencontré de pareils cas d'antéposition lors de leurs entraînements.

Enfin, dans certains cas, la multiplicité des erreurs rend la correction plus difficile. C'est le cas du couple de transcriptions tronquées ci-dessous. Le sujet syntaxique de *soumit* est ici séparé du verbe par plus de dix mots. De plus, la plateforme a sectionné l'énoncé en plusieurs phrases et a mal identifié plusieurs unités lexicales. Dans de telles situations, il est plus difficile d'envisager une amélioration du texte à l'aide d'un post-traitement.

REF	→	**RES**
Il ne monta rien, mais ravala (…) soumit	→	Il ne montre. Rien. Mais renvoie la (…) soumis

Intéressons-nous désormais aux erreurs pour lesquelles des post-traitements, s'ils ne sont pas nécessairement impossibles à mettre en œuvre, seront dans tous les cas plus difficiles. La Figure VI-14 ci-dessous montre que pour ce sous-corpus la majorité des erreurs relève de la substitution avec proximité phonétique. Pour quatre instruments, c'est d'ailleurs le seul type de substitutions rencontré. La proportion d'erreurs qui correspondent à des mots qui semblent avoir été ajoutés arbitrairement par les plateformes est très basse, allant de 0,15 % à 1,37 % des mots des fichiers RES.

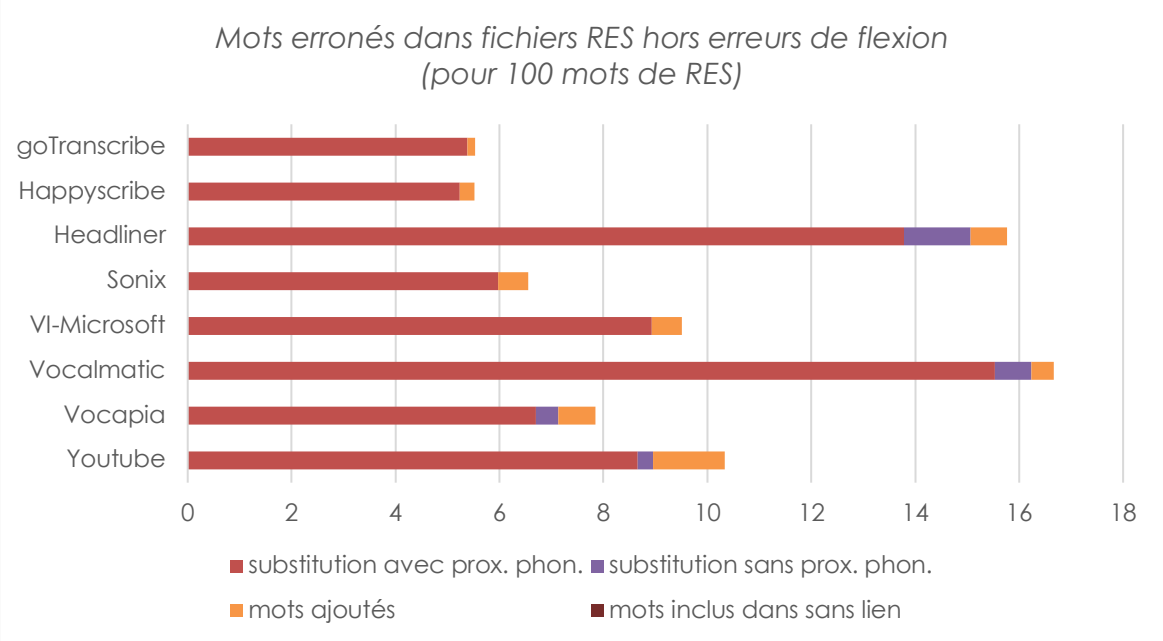


Figure VI-14 Répartition des erreurs hors erreurs de flexion par fichier RES — Physionomie

[30] On comptabilise trois *connues action*, trois *connu action* et 2 *connu actions*.

Pour compléter ces données chiffrées, observons quelques cas de substitutions :

REF	→	RES
Sous une *si vile* forme, *nous n'eussions*	→	sous une *ville* forme *nonu sion*
et *se manient à bonds*	→	et *semanie ah bon*
Cettui-ci	→	*Celui-ci*
hommes qui *furent onques*	→	homme qui *fut roncq*
prissassent	→	*prit sas*
et *se manient à bonds*	→	et *ce Mania bon*

Nous constatons qu'elles sont majoritairement liées à l'usage de temps verbaux propres à l'écrit comme le passé simple ou le passé antérieur (*n'eussions*, *saurions*, *prissassent*), ainsi qu'à la présence d'unités lexicales ou expressions inusitées (*cettui*, *se manient à bond*, *onques*). Comme nous l'avons présenté dans la section IV, plusieurs plateformes offrent la possibilité d'ajouter un lexique personnel avant de réaliser une transcription. Il serait intéressant de vérifier si l'ajout des mots inusités qui ont posé problème améliore la qualité des transcriptions obtenues.

Pour résumer, au-delà des bonnes performances observées pour chacune des plateformes pour la transcription du sous-corpus `Physionomie`, une marge d'amélioration automatique semble possible. Par ailleurs, la plateforme YouTube, même si elle a transcrit une plus faible proportion du signal, ne se démarque pas par un taux d'erreurs inférieur ni par l'absence d'un type particulier d'erreurs.

Comptines

Le sous-corpus `Comptines` apparaît comme étant le plus simple à transcrire pour l'ensemble des plateformes étudiées. Comme le montre la Figure VI-15, nous observons que le contenu des différents fichiers RES retournés par les plateformes est très proche du fichier de référence, avec une proportion moyenne de mots identiques à REF de 89,85 %. Dans le cas de cinq plateformes, plus de 90 % des mots de la transcription produite correspondent à des mots transcrits manuellement. L'écart entre les plateformes est un peu moins important que pour le sous-corpus `Physionomie`. Alors qu'il était de 13,45 %, il est ici de 10,33 %.

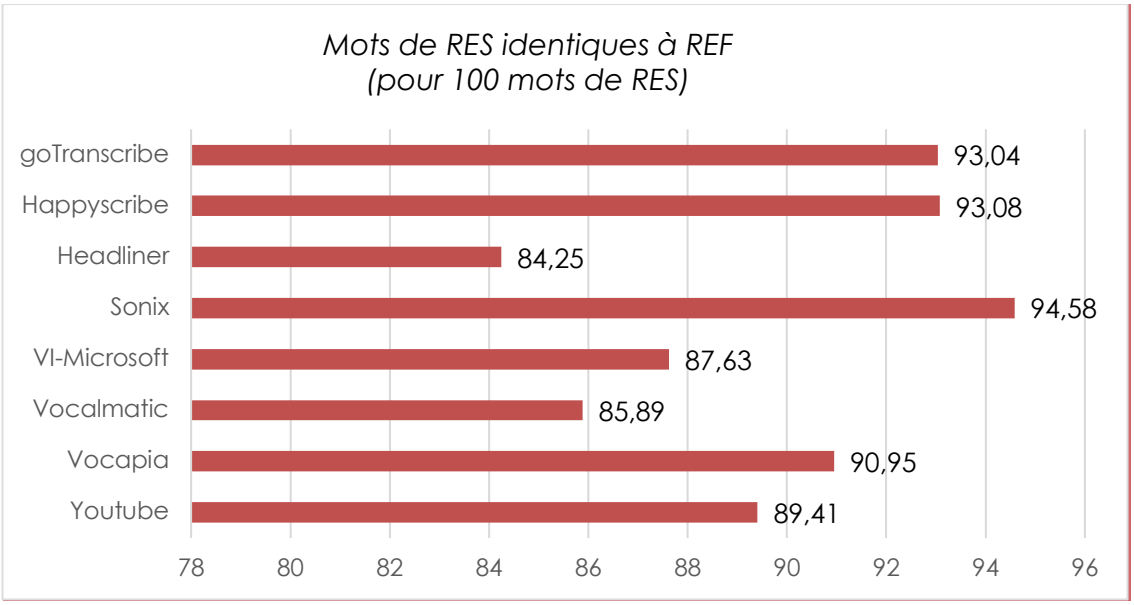


Figure VI-15 Proportion de mots identiques à REF dans chaque fichier RES — Comptines

Cette amélioration globale doit cependant être nuancée à la lecture de la Figure VI-16. Si YouTube propose ici une transcription plus couvrante que pour `Physionomie`, ce n'est pas le cas des autres instruments. La couverture reste toutefois tout à fait honorable et chaque instrument a proposé une transcription pour au moins 96 % des mots.

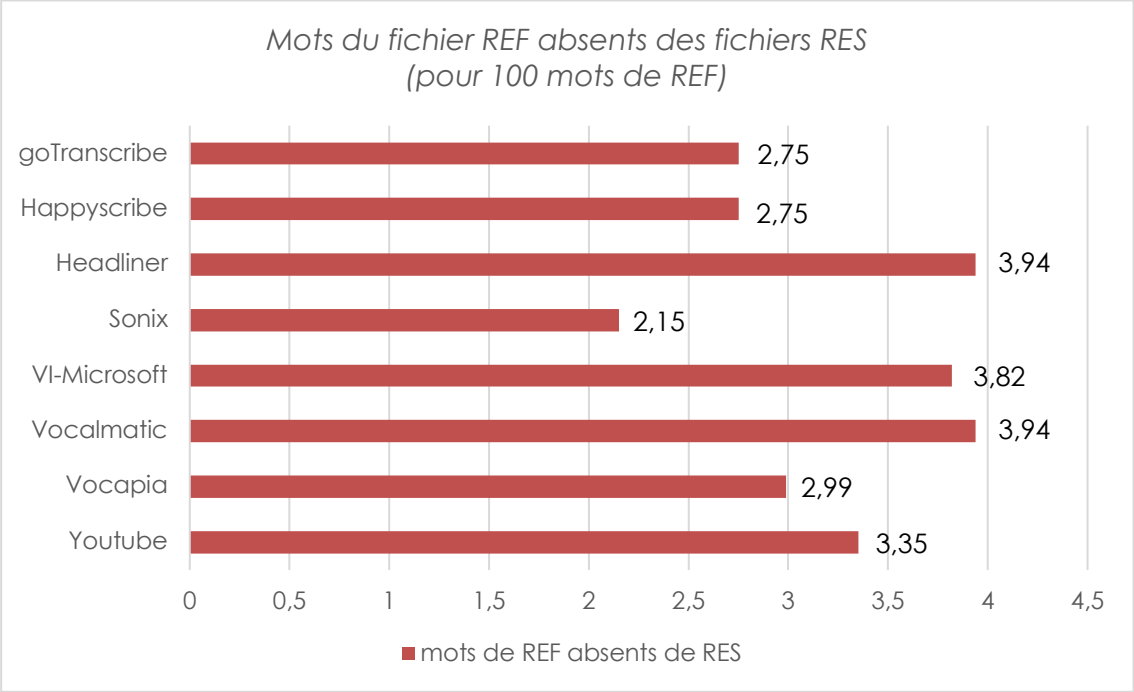


Figure VI-16 Proportion de mots de REF considérés comme n'ayant pas été transcrits — Comptines

En observant plus avant ces mots absents, on constate qu'ils relèvent pour beaucoup d'un choix éditorial que nous avons déjà évoqué. De la même façon que les

plateformes testées tendent à effacer les onomatopées et les interjections, elles semblent ignorer volontairement les répétitions de mots grammaticaux :

REF	→	RES
hein et puis *euh*	→	Et puis
on va rapprocher *des des* images *de de*	→	on va rapprocher *des* images
et et et les ressent	→	*et* les ressent

La Figure VI-17 présente les résultats que l'on obtient en procédant par ajout des mots identifiés comme comportant une erreur de flexion aux nombres de mots des fichiers RES identiques aux mots du fichier REF. Les erreurs de flexion sont ici moins nombreuses que dans les transcriptions produites pour le sous-corpus `Physionomie`. Elles représentent entre 2,27 % et 4,56 % des textes produits, pour une valeur médiane de 3,75 %. Leur correction automatique demeure une piste intéressante.

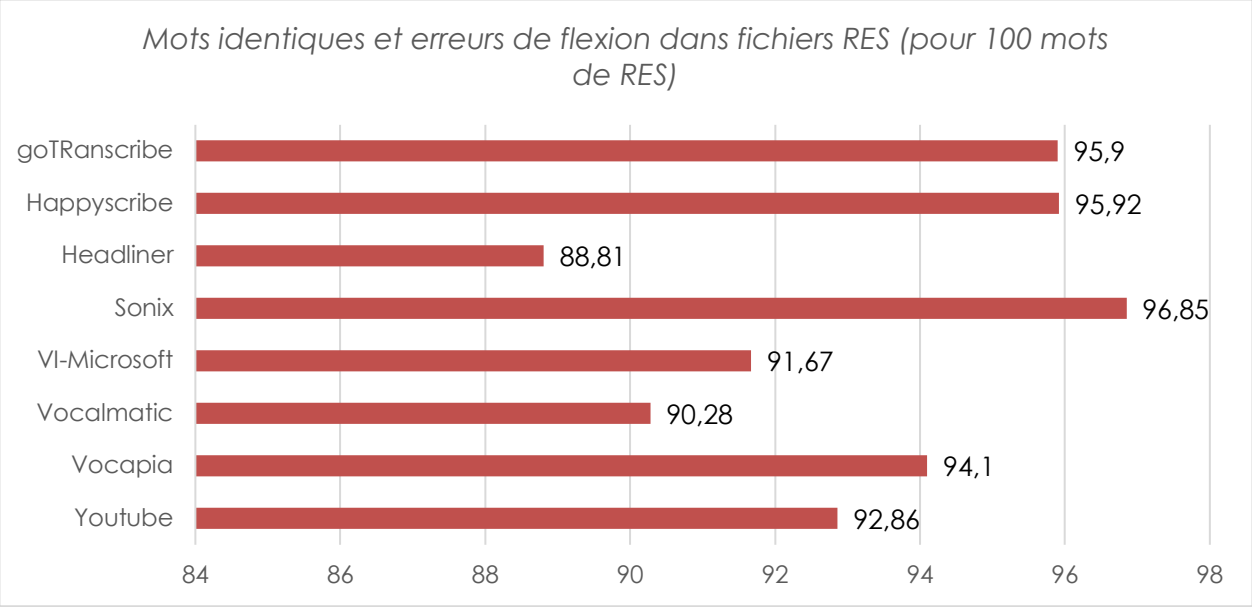


Figure VI-17 Proportion de mots identiques à REF ou ne variant que par la flexion dans chaque fichier RES — Comptines

En regardant plus en détail les erreurs de flexion rencontrées, on retrouve en partie des relations syntaxiques très simples à identifier :

REF	→	RES
qui sont tout à fait *admises*	→	qui sont tout à fait *admise*
certaines d'entre vous sont *institutrices*	→	certaines d'entre vous sont *institutrice*
leurs habitudes *ludiques*	→	leurs habitudes *ludique*
des tricheries *collectives*	→	des tricheries *collective*

Certaines autres nécessitent simplement un post-traitement qui prend en paramètre le genre de chaque locuteur :

REF	→	RES
je suis *frappée*	→	je suis *frappé*

On constate en revanche que d'autres sont plus difficiles à identifier. Elles sont associées à d'autres erreurs, comme le premier cas ci-dessous ou bien nécessitent une compréhension fine du propos pour être identifiées, comme les deux suivantes :

REF	→	RES
un peu plus *grands* ils découvriront	→	un peu plus *grand* et découvriront
à travers la notion d'*espace*	→	à travers la notion d'*espaces*
les enfants vont sans cesse dans *leurs aires* de *jeux*	→	les enfants vont sans cesse dans *leur aire* de *jeu*

De telles erreurs nécessitent une relecture manuelle soigneuse pour être corrigées. Elles viennent donc s'ajouter aux catégories d'erreurs pour lesquelles nous estimons qu'un post-traitement est aujourd'hui encore délicat à mettre en œuvre. La Figure VI-18 montre la répartition de ces erreurs dans les différentes transcriptions produites pour `Comptines`. On constate que leur proportion est plus basse que pour `Physionomie` pour l'ensemble des plateformes. Cependant, elles ont ici toutes produit des substitutions qui nous ont semblé sans proximité phonétique avec les mots de la transcription de référence. La quantité d'erreurs qui correspondent à des mots qui semblent avoir été ajoutés arbitrairement est un peu plus élevée que pour `Physionomie`, mais demeure très basse, allant de 0,25 % à 1,77 % des mots des fichiers RES.

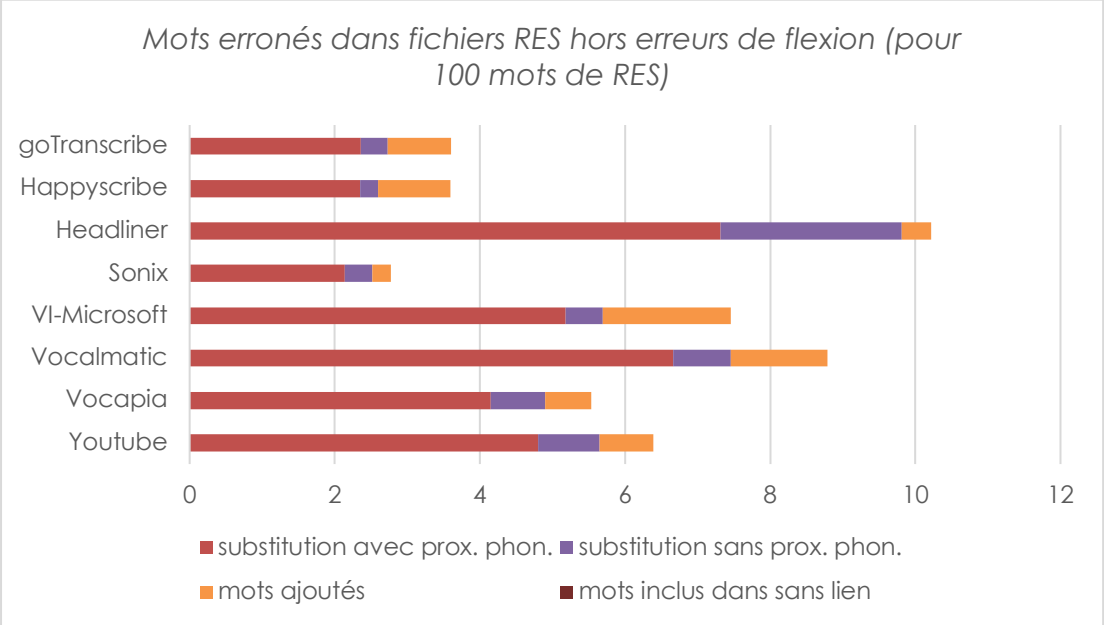


Figure VI-18 Répartition des erreurs hors erreurs de flexion par fichier RES — Comptines

Observons ici aussi quelques cas de substitutions avec proximité phonétiques :

REF	→	RES
sans qu'on leur *en ait*	→	*sont* qu'on leur *amenait*
va prendre	→	*d'apprendre*
des *rythmes*	→	des *rites*
turlututu	→	*sur le tutu*
qui *s'écoule*	→	qui *c'est cool*
des nombres et	→	*dénombrer*

Contrairement à ce que nous avions observé pour `Physionomie`, ces erreurs ne relèvent pas de spécificités du genre du sous-corpus ou du registre de langue employé. À part *turlutu*, que l'on peut supposer absent des lexiques exploités par les instruments, aucune des formes substituées n'est particulièrement rare ni ne relève d'un vocabulaire spécialisé.

Il en va de même pour les substitutions sans proximité phonétiques :

REF	→	RES
d'un coup	→	*d'abord*
qui *l'est*	→	qui *finit*
il *y a*	→	il *vient*
qu'on	→	*pour*

En conclusion, les transcriptions du sous-corpus `Comptines` proposées par les différentes plateformes testées sont de bonne qualité. Cependant, un certain nombre de phénomènes en ont été exclus (onomatopées, interjections, répétitions) et la marge de manœuvre pour améliorer ces performances semble limitée. Ni l'ajout de lexique en entrée des plateformes ni l'intégration de post-traitements syntaxiques ne nous semblent ici pertinents.

Camille

On peut avoir le sentiment dans un premier temps, à la lecture de la Figure VI-19, que les transcriptions proposées pas les différentes plateformes pour le sous-corpus `Camille` sont d'une qualité proche de celles proposées pour `Physionomie`. La proportion de mots des fichiers RES qui se trouvent à l'identique dans le fichier REF est même légèrement supérieure ici (située entre 78,17 % et 86,61 %). À l'exception de Video Indexer, qui chute en dernière place du classement, les différents instruments continuent de se situer à peu près de la même manière les uns par rapport aux autres.

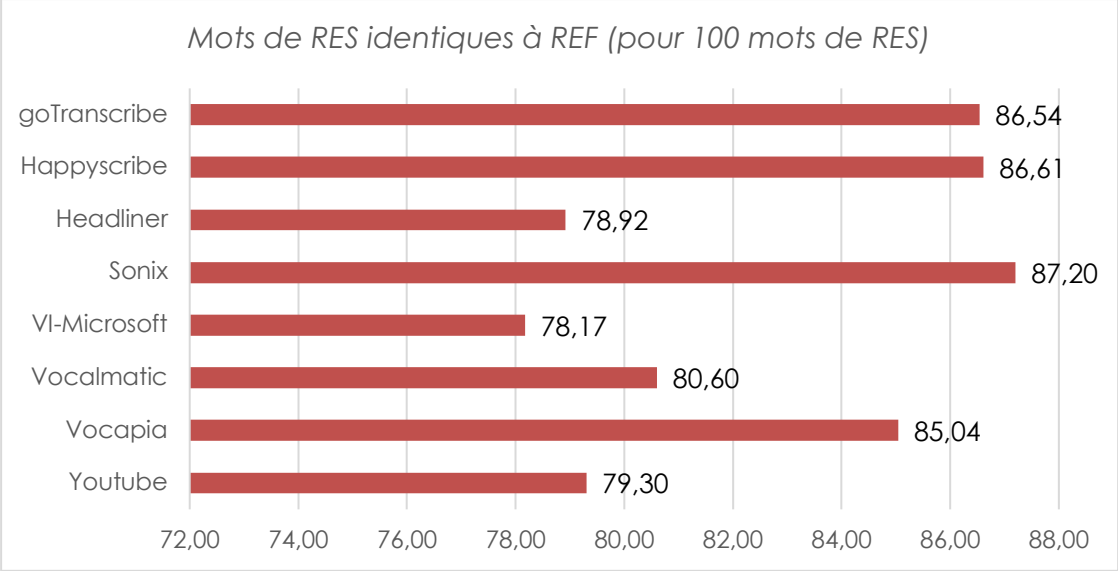


Figure VI-19 Proportion de mots identiques à REF dans chaque fichier RES — Camille

Attention ce sentiment est cependant vite détrompé lorsque l'on s'intéresse à la taille des transcriptions produites et, par conséquent, à la proportion des mots de REF pour lesquels nous avons considéré que les différents instruments n'avaient rien proposé du tout. La Figure VI-20 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 96 % des mots.

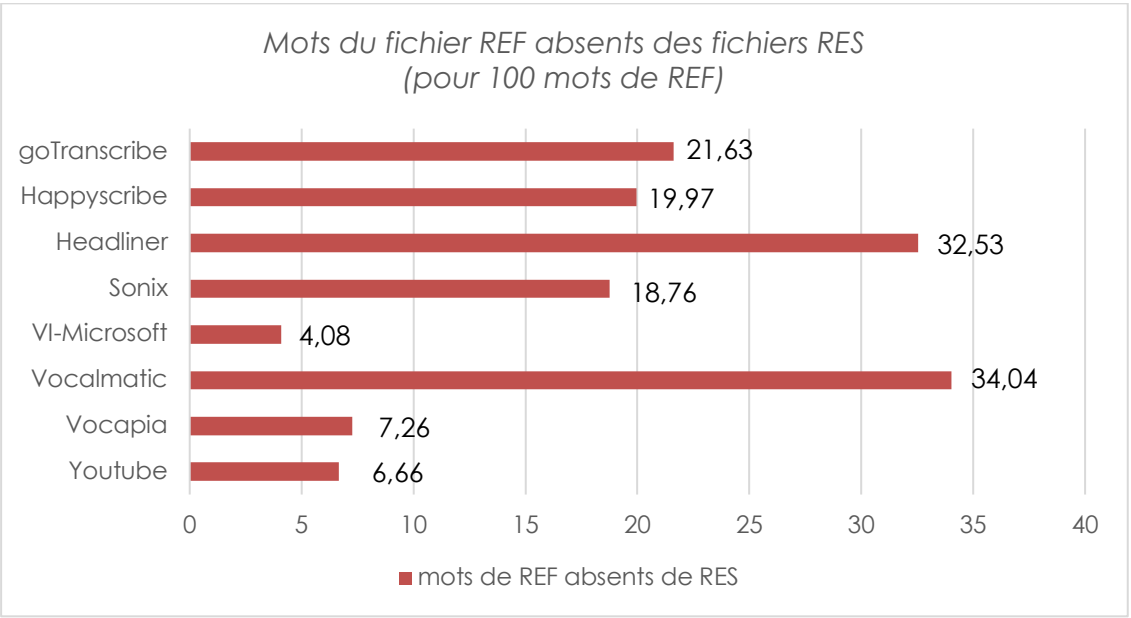


Figure VI-20 Proportion de mots de REF considérés comme n'ayant pas été transcrits — Camille

Nous chercherons par la suite à évaluer si les trois groupes qui se dessinent correspondent à trois stratégies différentes face à des passages difficiles à transcrire. Mais avant cela, intéressons-nous aux erreurs de flexion rencontrées. Comme le montre la Figure VI-21, elles sont encore moins nombreuses que pour le sous-corpus

`Comptines`. Leur correction exhaustive représenterait une amélioration des textes RES de 1,18 % à 2,44 %.

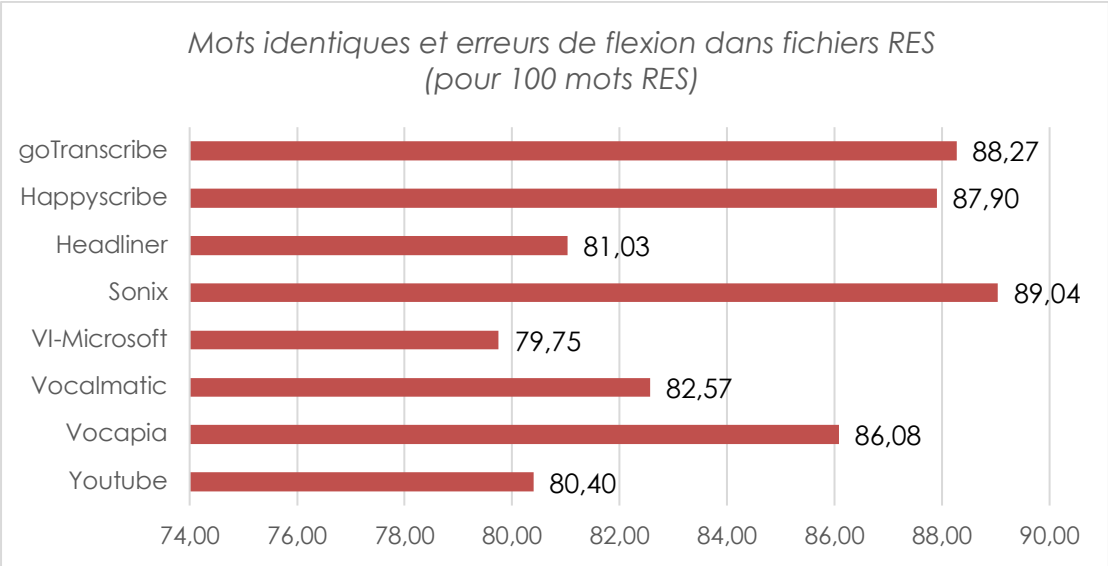


Figure VI-21 Proportion de mots identiques à REF ou ne variant que par la flexion dans chaque fichier RES — Camille

En regardant en détail les erreurs de flexion rencontrées, on trouve majoritairement des cas relativement difficiles à identifier. Leur correction nécessiterait une compréhension fine du propos, comme la première ci-dessous, ou une réécoute de la version audio du sous-corpus, comme les deux suivantes :

REF	→	**RES**
des études de *dossiers*	→	des études de *dossier*
et je on *fait* de l'accompagnement	→	et je vous *fais* de l'accompagnement
ça *sonnera* dans 20 mn.	→	ça *sonne* dans 20 minutes

Les cas qu'il semble possible de corriger à l'aide d'un post-traitement automatique sont rares, même si nous en avons rencontré quelques-uns :

REF	→	**RES**
est-ce que tu me *raconterais*	→	est-ce que tu me *raconterai*
association qu'il avait *eue*	→	association qu'il avait *eu*

Les autres catégories d'erreurs, comme le montre la Figure VI-22, représentent en moyenne une plus grande proportion de mots que ce que nous avons observé dans les deux autres sous-corpus (13,83 %, contre 9,71 % pour `Physionomie` et 6,04 % pour `Comptines`). De plus, les fichiers RES annotés ici sont les seuls à comporter des passages dits « sans lien ».

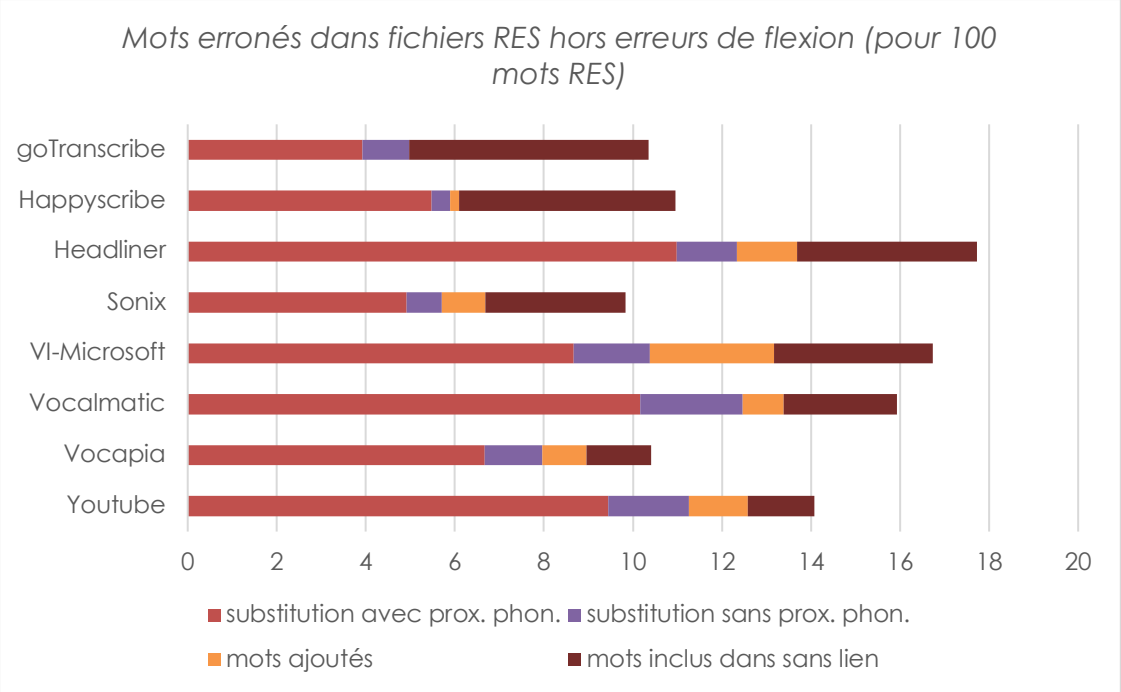


Figure VI-22 Répartition des erreurs hors erreurs de flexion par fichier RES — Camille

La catégorie d'erreurs majoritairement rencontrée ici n'est plus clairement la substitution avec proximité phonétique pour l'ensemble des plateformes. Trois d'entre elles se distinguent. Elles correspondent à l'un des groupes que nous avions évoqués tout à l'heure : Sonix, Happy Scribe et Go Transcribe. Rappelons-le, les textes qu'elles ont produits ne sont pas les plus couvrants de tous, il leur manque environ 20 % des mots transcrits manuellement. En revanche, les mots contenus dans leurs fichiers RES sont à plus de 86 % identiques à ceux du fichier REF. Ce que l'on observe ici c'est que ces instruments ont introduit peu ou pas de mots qui nous semblent avoir été ajoutés arbitrairement, une faible quantité de substitutions sans proximité phonétique et les plus basses proportions de substitutions avec proximité phonétique. En revanche ils retournent un nombre important de passages (respectivement 8, 8 et 10) que nous avons jugés sans lien avec la transcription manuelle de référence. Ainsi le nombre de mots impliqués dans ces passages est très proche du nombre de ceux impliqués dans des substitutions avec proximités phonétiques.

Le sous-corpus `Camille` comporte davantage d'entités nommées que les deux autres étudiés. Nous avons pensé dans un premier temps que ces entités nommées étaient inconnues des plateformes, ce qui était à l'origine de nombreuses substitutions. Nous avons effectivement observé de telles substitutions[31]. Cependant, en regardant plus en détail, nous avons constaté qu'à l'exception de *Carnin*, ces entités nommées sont toutes transcrites correctement par au moins une plateforme. De plus, lorsque le sous-corpus comporte plusieurs occurrences d'une même entité nommée, la plupart des plateformes transcrivent correctement au moins une de ces occurrences.

[31] Pour préserver l'anonymat, nous ne pouvons pas illustrer ces substitutions.

Les autres erreurs de substitution rencontrées, avec ou sans proximité phonétique, quant à elles, ne relèvent pas d'un lexique spécifique :

REF	→	RES
puisqu'on	→	*parce qu'on*
dans 20 min	→	*très vite*
ce que t'es	→	*parce que*
j'ai *en charge*	→	j'ai *un chargeur*
à ces	→	*assez*

L'ajout de lexique en entrée ne nous semble donc pas, dans ce cas précis, une piste à privilégier pour améliorer les performances.

Reprenons à présent la question des trois groupes et de l'existence de stratégies différentes face à des passages difficiles à transcrire. Pour cela, nous avons choisi de comparer les proportions de mots de REF absents des textes produits par les différentes plateformes : à la fois ceux que nous avons catégorisés comme mots de REF absents de RES et ceux comptabilisés dans les nombres de mots de REF impliqués dans des passages sans lien. La Figure VI-23 montre le résultat de cette comparaison.

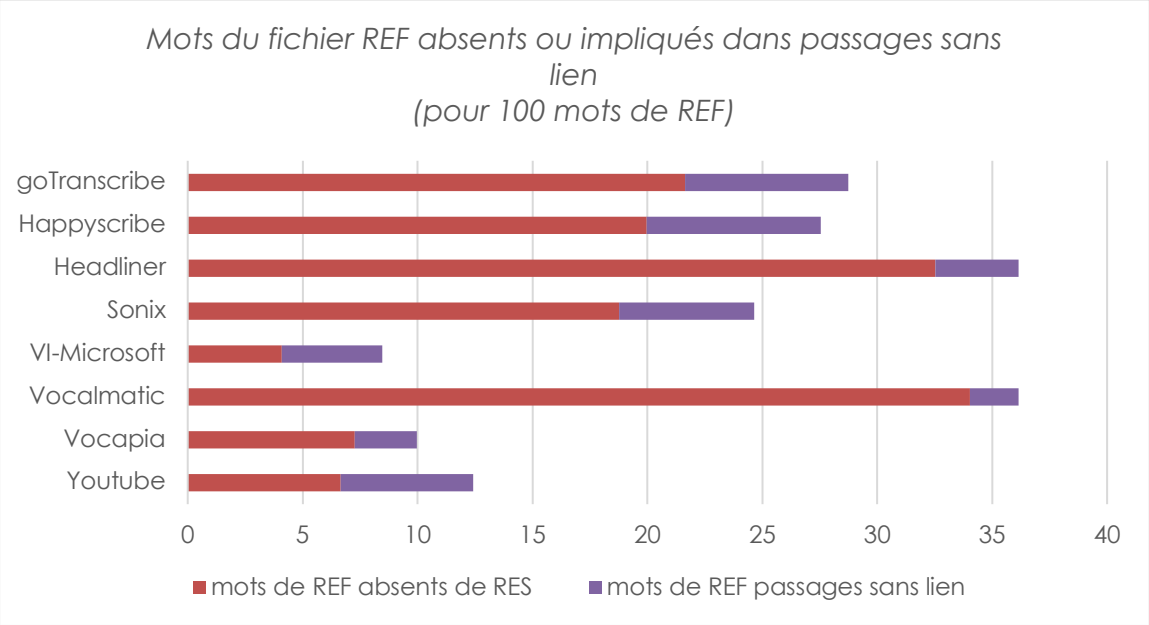


Figure VI-23 Proportion de mots de REF considérés non transcrits ou impliqués dans des passages sans lien — Camille

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). 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). 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), la transcription proposée par Vocapia n'en comporte que 6, relativement courts. 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 `Camille` 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 sortir son épingle du jeu, en proposant une transcription à la fois couvrante et d'assez bonne qualité.

Conclusions

Nous avons pu voir tout au long de notre analyse que 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 `Camille`. Elles ne sont pas nécessairement suffisantes pour répondre à tous les cas de figure et le lecteur est invité à ajouter ses propres catégories selon ses besoins et la nature de ses observations.

VII. Estimation du gain de temps

Le dernier volet de l'analyse des résultats a consisté à estimer la qualité des textes obtenus à partir d'une lecture subjective de ces textes, ainsi que d'une estimation des gains de temps réalisés en utilisant ces outils, par rapport à une transcription manuelle intégrale. Y a-t-il correspondance entre les scores WER obtenus, le type d'erreur repérées et le temps mis à corriger ?

Lecture qualitative des textes retranscrits

L'objectif de cette lecture était de vérifier si — malgré les erreurs de transcription automatique — le texte obtenu était utilisable par le professionnel qui va l'exploiter. Le niveau d'exigence étant, bien entendu, différent selon l'usage final prévu.

Le choix a été fait de partir des fichiers bruts obtenus par l'utilisateur. Certains nécessitent en effet d'effectuer une première manipulation, potentiellement chronophage, pour obtenir un format exploitable en lecture humaine (fichier texte brut .TXT ou fichier Word .DOC). Quelles informations (objectives et subjectives) repère-t-on à la lecture de ces fichiers ?

Le point commun observé sur l'ensemble des fichiers est la nécessité de corriger les nombreuses fautes d'orthographe et de grammaire qui émaillent les transcriptions.

Pour `Physionomie`, aucune transcription n'est parvenue à décoder *Cettui-ci* qui devient à chaque fois *celui-ci*. Globalement les termes en vieux français n'ont été que très rarement identifiés (*onque*, *prissassent*…).

Comme cela a été signalé auparavant, toute une série d'erreurs relève de problèmes d'homophonie et force est de constater que le choix orthographique fait par la plateforme tombe souvent à côté…

Les limites de l'exercice avec `Harmonie`. Ce fichier se caractérise par une prise de son qui a capté les bruits environnants (chuchotements, bruits parasites) dans le cadre d'une réunion où les interlocuteurs sont au-delà de quatre et parlent quelquefois en même temps. La transcription du fichier audio s'avère difficile même pour une personne aguerrie à ce genre d'exercice. Certains passages du fichier audio ne peuvent pas être retranscrits soit parce que plusieurs personnes parlent en même temps, soit parce que l'écoute est parasitée par des bruits divers (manipulation de feuilles de papier et/ou objets présents sur la table), ou bien encore parce que l'interlocuteur est trop loin du micro. Ce fichier nous rappelle combien il est important de disposer d'un équipement adapté à nos besoins lorsque l'on se livre à des enregistrements et en particulier lorsqu'il y a plusieurs interlocuteurs.

Des similitudes d'erreurs entre plateformes. Les plateformes Happy Scribe et Go Transcribe « commettent » le même type d'erreurs. Par exemple *c'est-à-dire qu'ils savent toujours sur qui ça va finir* est devenu pour ces deux plateformes *c'est-à-dire qu'ils sortent toujours en Turquie ça va finir*. Certaines erreurs sont communes à trois plateformes (Happy Scribe, Sonix et Go Transcribe) ; exemple : *constituer socialement leur dossier MDPH* est devenu *constituer en Bosnie MDPH*.

Les catégories utilisées pour classer les plateformes sont présentées dans le Tableau VII-1 et les résultats dans le Tableau VII-2.

Bonne transcription	Peu d'erreurs de vocabulaire ou de syntaxe dans le fichier	Ce sont essentiellement des erreurs d'homophonie que l'on rencontre : *cocher/kosher* ; *vaine/veine*.
Texte utilisable	Les erreurs majoritaires sont en lien avec l'homophonie	Par exemple la transformation du mot *comptine* en *cantine*, ou bien *les chiffres n'ont pas de fin* par *les chiffres n'ont pas de faim* ; erreurs d'accord en lien avec l'homophonie entre un pluriel et un singulier : *Celles qui coulent sous la naïveté/Celle qui coule sous la naïveté*
Travail important de correction à prévoir	Plusieurs phrases sont ineptes et nécessitent une réécriture à l'écoute	*Je t'appellerai Camille* est devenu *je t'appellerai terrier e gaming* ; *ce mystère de l'infini* est devenu *ce mystère de la fille* ou bien encore *chacun l'entend* est devenu *Shaka longtemps*
Transcription inutilisable	Peu de matériau utilisable pour simplifier le travail du relecteur	À titre d'exemple, on oscille entre le cadavre exquis : *lui par le passé avec déjà des bleus, mais* ; ou une réplique de maître Yoda sous extasie : *Raconterai excuse-moi, je vais mettre*.

Tableau VII-1 Catégories pour l'estimation qualitative des textes transcrits

	Happy Scribe	Go Transcribe	Sonix	Vocapia	Video Indexer	YouTube	Headliner	Vocalmatic
Physionomie								
Comptines								
Camille								
Harmonie*								

Tableau VII-2 Estimation qualitative de la qualité des textes retranscrits.
Vert : bonne transcription ; Violet : texte utilisable ;
Orange : travail important de correction à prévoir ; Rouge : transcription inutilisable
*Le fichier de base étant complexe, la comparaison n'a pas été réalisée

Un premier constat est que 80% des fichiers sont considérés comme exploitables. Sur les trois fichiers audio transcrits (`Camille` — `Comptines` — `Physionomie`), `Comptines` est le fichier audio pour lequel les résultats sont les meilleurs. Cela rejoint les résultats obtenus dans les deux sections précédentes.

Cependant, il y a apparemment des contradictions entre scores du WER et l'évaluation subjective réalisée ici. La différence est particulièrement saillante pour `Physionomie` obtenu par Go Transcribe : bien qu'obtenant un des meilleurs WER, le

fichier est estimé carrément inutilisable ici. À l'inverse, `Comptines` obtenu par Headliner pourrait être plus intéressant que les résultats du WER ne le laissent penser.

Afin de vérifier ces premiers résultats, un calcul de gain de temps a été effectué par l'une des membres du groupe pour tous les fichiers.

Correction sans mise en forme

Le calcul des gains de temps a été réalisé par une chargée d'études qui travaille habituellement sur des études et recherches en lien avec la relation formation/emploi. Selon les travaux menés, ce sont 150 à 250 entretiens — d'une durée de 1 h à 2 h — qui vont être réalisés. L'objectif attendu des transcriptions est d'avoir un texte qui rende compte de l'entretien afin que leur exploitation permette de renseigner des hypothèses et indicateurs nécessaires à l'analyse d'une problématique. Il ne s'agit pas d'être dans une transcription mot à mot, pour autant, il s'agit de rester fidèle au discours obtenu. Dans cette logique, les marques d'énonciation (répétitions…) ne sont pas corrigées, mais les hésitations ou silences ne sont pas pris en compte (Tableau VII-3). Le principe appliqué est celui de reconstituer des phrases qui font sens (enlever les mots ineptes, ajouter ceux qui manquent) et de rétablir l'orthographe.

Transcription mot à mot (Camille)	**Transcription pour l'exercice** (Camille)
L : D'accord. Tu vois combien de personnes en moyenne qui viennent à ces permanences ?	L : D'accord. Tu vois combien de personnes en moyenne qui viennent à ces permanences ?
C : **Euh…** en moyenne je sais que par exemple le mercredi matin on a… alors nous il faut savoir qu'on prend le temps d'expliquer aux gens, on prend le temps de bien poser, de bien expliquer, s'assurer que la personne à compris ce qu'elle … donc ce sont des entretiens qui peuvent prendre du temps déjà. **Euh… et ensuite ben…** ensuite en moyenne on met, on a deux rendez-vous par matinée je sais, le mercredi en moyenne, parfois trois, c'est un, c'est un peu ric-rac, ça dépend, ça dépend **euh…** l'importance de la demande.	C : En moyenne je sais que par exemple le mercredi matin on a… alors nous il faut savoir qu'on prend le temps d'expliquer aux gens, on prend le temps de bien poser, de bien expliquer, s'assurer que la personne à compris ce qu'elle… donc ce sont des entretiens qui peuvent prendre du temps déjà. Et ensuite en moyenne on met, on a deux rendez-vous par matinée je sais, le mercredi en moyenne, parfois trois, c'est un, c'est un peu ric-rac, ça dépend, ça dépend de l'importance de la demande.

Tableau VII-3 Comparaison d'une transcription mot à mot avec la transcription conservée pour l'exercice

Le calcul des gains de temps 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". Pour chaque fichier le temps de transcription des 2'30" est le suivant[32] :

- Physionomie : 14' ou 840"
- Comptines : 20' ou 1200"
- Camille : 18' ou 1080"
- Harmonie : 22' ou 1320"

Ces temps ont été utilisés comme base pour le calcul des gains possibles en utilisant les différentes plateformes.

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 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 `Comptines` et la version donnée par YouTube), nécessitent un temps de correction supérieur à ce qui a été évalué de façon qualitative (relecture par un humain).

Ce premier travail met en évidence une dissonance entre l'évaluation faite par un relecteur des fichiers obtenus en mode brut harmonisé sur le temps que nécessiterait la remise en conformité du texte.

La représentation subjective du temps à passer pour corriger les transcriptions obtenues à partir de la lecture des fichiers, ne coïncide pas avec le temps réellement passé pour corriger les fichiers. Le Tableau VII-4 représente ce décalage.

	Happy Scribe	Go Transcribe	Sonix	Vocapia	Video Indexer	YouTube	Headliner	Vocalmatic
Physionomie	7'20"	7'54"	7'12"	8'16"	8'13"	11'26"	8'24"	10'51"
Comptines	5'53"	5'57"	5'05"	6'20"	7'32"	8'03"*	9'00"	10'29"
Camille	8'26"	9'09"	8'16"	6'14"	6'40"	9'11"	11'00"	11'07"
Harmonie*	12'37"	15'15"	15'02"	11'28"	12'32"	12'46"	16'08"	16'21"

Tableau VII-4 Durée de correction du texte seul. Le code couleur représente les estimations précédemment faites à la lecture du texte retranscrit. Vert : bonne transcription, violet : texte utilisable, orange : travail important de correction à prévoir, rouge : transcription inutilisable

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 VII-5), cela modifie sensiblement les résultats pour certaines plateformes (Tableau VII-6).

[32] La personne ayant réalisé ce travail n'est pas une secrétaire confirmée, mais elle est habituée à retranscrire des entretiens et utilise ses dix doigts sur le clavier et peut atteindre 43 mots minutes avec une précision de 86,81 % (Test réalisé sur http://typing-test.arckade.fr/).

	Happy Scribe	Go Transcribe	Sonix	Vocapia	Video Indexer	YouTube	Headliner	Vocalmatic
Physionomie	0'16"	0'17"	0'22"	1'07"	01'41"	2'32"	2'34"	0'39"
Comptines	0'27"	0'20"	0'17"	0'12"	2'57"	2'46"	2'23	0'08"
Camille	0'44"	0'00"	0'19"	0'33"	2'47"	2'04"	3'19"	0'17"
Harmonie	0'47"	0'15"	01'31"	0'28"	2'06"	1'35	1'18"	0'23"

Tableau VII-5 Durée de correction de la mise en forme du texte

	Happy Scribe	Go Transcribe	Sonix	Vocapia	Video Indexer	YouTube	Headliner	Vocalmatic
Physionomie	7'36"	8'11"	7'34"	9'23"	9'54"	13'58"	10'58"	11'30"
Comptines	8'16"	6'24"	5'25"	6'32"	10'23"	10'49"	9'17"	10'37"
Camille	9'10"	9'9"	8'35"	6'47"	8'27"	11'15"	14'19"	11'24"
Harmonie	13'24"	15'30"	16'33"	11'56"	14'38"	14'21"	17'26"	16'44"

Tableau VII-6 Durée cumulée de la correction du texte et de la mise en forme

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 :

- *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,
- Le temps d'édition par la plateforme des données (temps d'obtention des fichiers).

Le tableau VII-7 présente les gains de temps en % par plateforme et par corpus. Les figures VII-1 et VII-2 proposent une représentation sous forme de boite à moustache de ces valeurs.

	Go Transcribe	Happy Scribe	Headliner	Sonix	Video Indexer	Vocalmatic	Vocapia	Youtube
Physionomie	42%	46%	22%	46%	29%	18%	33%	0%
Comptines	68%	59%	54%	73%	48%	47%	67%	46%
Camille	49%	49%	20%	52%	53%	37%	62%	38%
Harmonie	30%	39%	21%	25%	33%	24%	46%	35%

Tableau VII-7 Gain de temps en % par plateforme et corpus

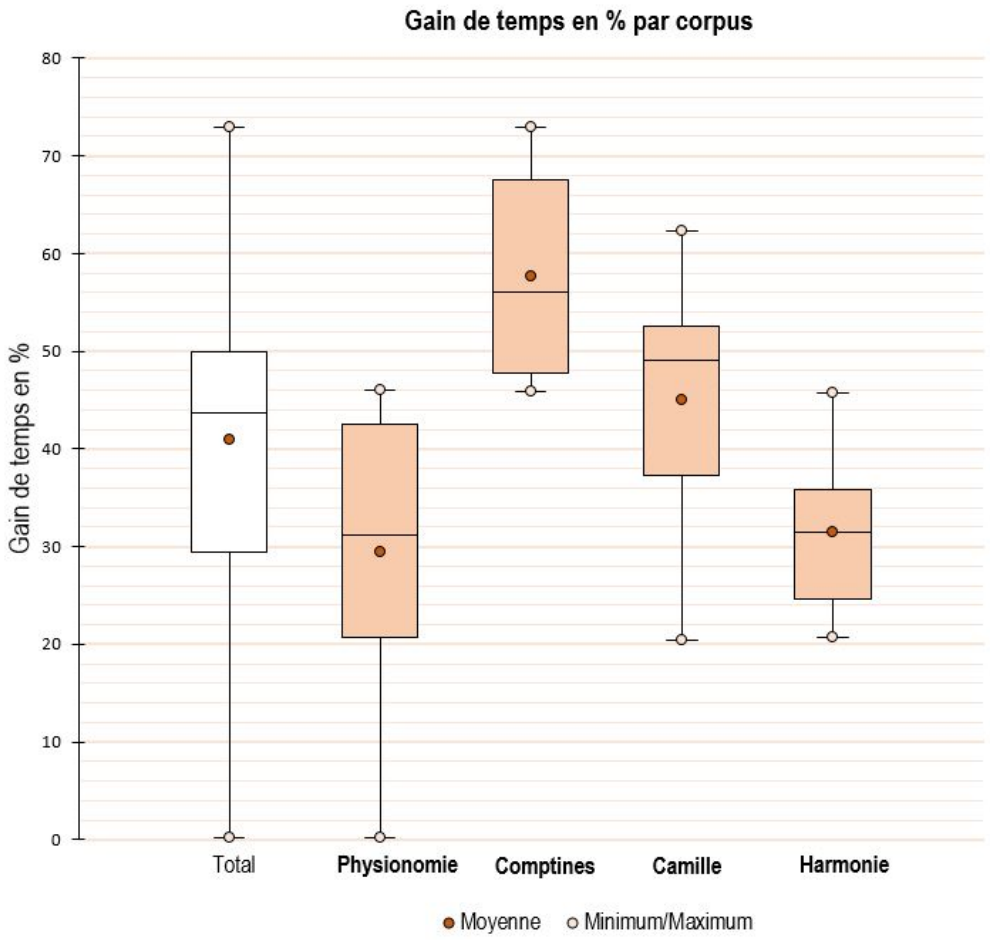


Figure VII-1 Gain de temps par corpus

Le gain de temps est variable et n'existe pas toujours. Quand il existe, il est à minima de 18 % et peut monter à 73 %. Sans surprise, c'est `Comptines,` le cours magistral, qui obtient le meilleur gain avec une médiane à 56 % de gain de temps. Même pour `Camille`, l'entretien sociologique avec une élocution problématique, le gain est notable (médiane à 49 %). `Physionomie` est tiré vers le bas par la (très) mauvaise prestation de YouTube sur ce fichier, qui déséquilibre la série. Enfin, sur `Harmonie`, 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. On remarque tout de même que le WER de `Physionomie` était globalement bien meilleur que celui d'`Harmonie` 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.

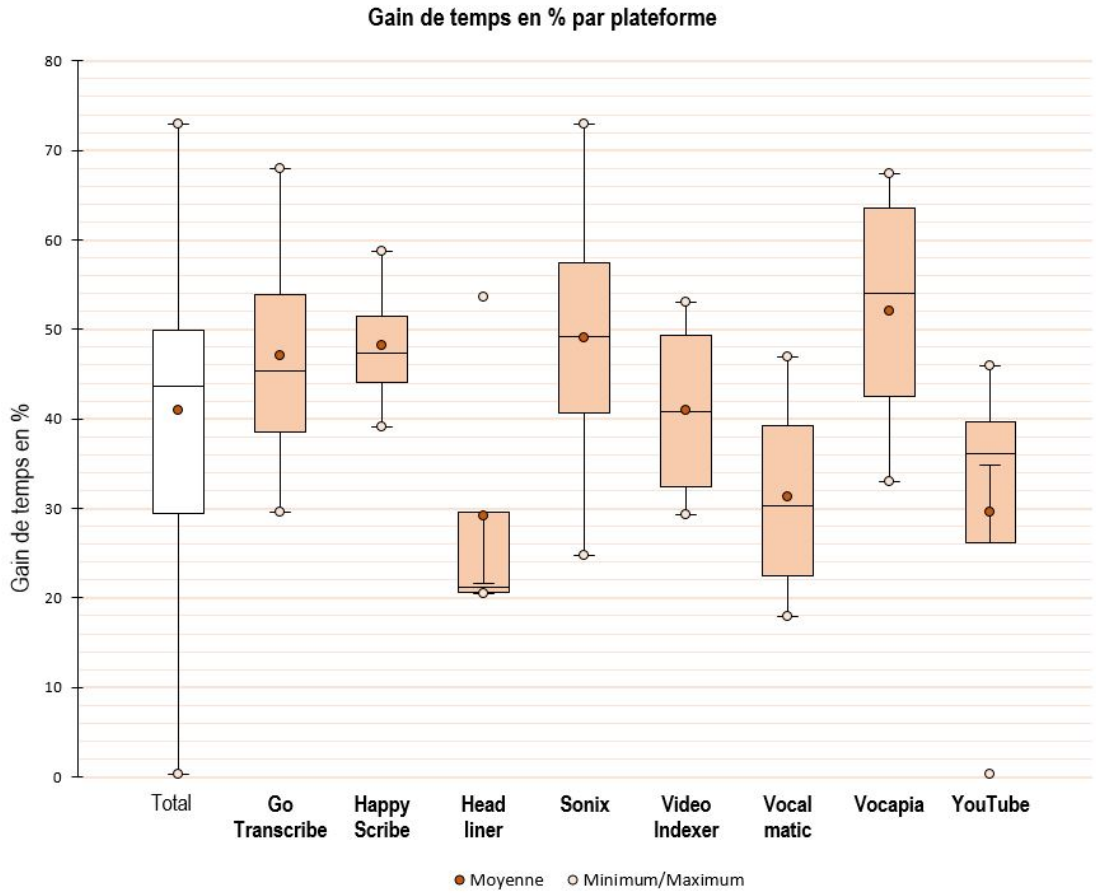


Figure VII-2 Gain de temps par plateforme

D'après la Figure VII-2, Headliner et Youtube se démarque, avec chacun 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, Physionomie-Happy Scribe obtient un WER de 15 et Camille-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.

Si nos comparaisons des résultats obtenus par les plateformes encouragent à utiliser ce genre d'outils dans une perspective purement économique (monétaire ou temps), il y a pourtant d'autres raisons qui pourraient pousser à s'éloigner de ce type d'outils. Nous allons les présenter après avoir abordé les limites et perspectives de cette étude.

VIII. Limites et perspectives

Un domaine en constante évolution

Le marché de la transcription automatique est en perpétuelle évolution, et ses changements sont très rapides. Nos tests ont été effectués entre novembre 2019 et février 2020. Dès à présent, et encore plus dans quelques mois, les résultats ne seront sans doute plus les mêmes. Il est toutefois difficile de prédire comment la situation va évoluer. Certains outils seront sans doute amenés à disparaître, quand d'autres, notamment institutionnels, peuvent être amenés à se développer. L'Organisation Mondiale de la Propriété Intellectuelle en est un exemple, avec l'outil qu'elle a conçu pour traiter des corpus de réunions et conférences internationales. Le logiciel, « créé à l'origine pour aider à transcrire les réunions officielles de l'OMPI, peut être personnalisé pour d'autres organisations »[33].

Ce que l'on aurait aimé faire, mais qui n'a pas été possible

Un certain nombre de points ont régulièrement été soulevés au cours de ce travail, et mériteraient d'être étudiés par la suite.

Le premier de ces points concerne la variabilité des situations enregistrées, que notre corpus ne permet pas de couvrir. Deux points nous semblent particulièrement importants à évaluer dans le futur : (1) la variation des dispositifs d'enregistrement (téléphone, dictaphone, smartphone, régie, etc.) et la diversité des locuteurs (par ex. en termes de genre, âge, locuteurs natifs ou non natifs, locuteurs avec différents accents régionaux, locuteurs avec différentes difficultés phonatoires).

Le second de ces points concerne l'évaluation de la qualité de la détection des changements de locuteurs. Celle-ci nécessiterait la mise en place d'une méthodologie adéquate, qui n'a pas pu être réalisée dans l'intervalle de temps que nous nous étions fixé. Comme le mentionnent Bazillon et al. (2008) : « certains types de données (parole téléphonique ou locuteurs non natifs par exemple) sont un peu problématiques pour les systèmes ASR, alors qu'elles ne perturbent pas un annotateur humain. L'affectation des locuteurs peut sembler assez longue pour la parole spontanée, mais ces résultats doivent être limités : généralement, la transcription du texte, l'affectation des locuteurs et la correction de l'orthographe se font au fur et à mesure, au lieu d'être séparées. Dans le cas des orateurs, cela est important, car leur assignation après la transcription oblige le transcripteur à vérifier l'ensemble du fichier. Dans le cas d'un discours préparé, cela ne prend pas trop de temps, car les tours de parole sont généralement longs et bien définis ; à l'inverse, le discours spontané contient souvent des tours de parole courts avec de nombreux changements de locuteurs ».

Un troisième point concerne l'ajout de lexiques personnalisés en amont de la transcription. Ceux-ci permettent une amélioration notable de la qualité de la

[33] https://www.wipo.int/about-ip/en/artificial_intelligence/speech_to_text.html

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.

Enfin, un quatrième et dernier point concerne l'analyse du fichier le plus problématique, celui de la réunion associative (`Harmonie`). Caractériser les erreurs obtenues pour ce fichier s'est avéré une vraie gageure, qui demandait là encore plus de ressources que ce dont nous disposions. Cela nécessiterait la mise en place d'une méthodologie adaptée à la réception de textes très fragmentaires.

Des raisons de ne pas utiliser ce genre d'outils

Un dernier point que nous souhaitions mentionner concerne une réflexion plus globale, portant sur le recours à ce genre d'outils. Bien que leurs promesses de gain de temps et d'argent soient en partie réalisées, il existe des arguments à la fois déontologiques, épistémologiques et politiques qui vont à l'encontre de leur usage.

Les arguments déontologiques ont été développés dans la section sur la confidentialité des données, nous ne reviendrons donc pas dessus. Ils restent cependant fondamentaux à prendre en compte au moment de faire le choix des outils utilisés, et plaident notamment pour le développement d'outils académiques, libres et hébergés sur des serveurs nationaux.

Des arguments épistémologiques vont aussi à l'encontre de l'usage de ce type d'approche. Comme le résume très bien Mondada (2000), « 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 ». Transcrire, c'est déjà analyser : déléguer ce travail peut être vu comme problématique dans un certain nombre de cas. Bien entendu, des pratiques de recherche bien établies font déjà appel à une forme de délégation, en employant par exemple des étudiants pour réaliser ce genre de travail. Une dimension formative est néanmoins à l'œuvre dans ce cas, qui sera transformée par ces outils. Les savoirs acquis et les réflexions menées sur le matériau ne seront pas les mêmes suivant que l'on effectue un travail de correction plutôt que de transcription intégrale.

L'automatisation d'un certain nombre de tâches ravive également le débat sur le lien entre automatisation et perte d'emplois. Il n'existe pas, à notre connaissance, d'enquête sociologique sur les travailleuses et travailleurs de la transcription. Il est donc difficile d'évaluer les conséquences que l'émergence de ce type de plateformes peut avoir sur cette profession (reconfiguration, disparition), et à quel prix. En outre, un certain nombre de plateformes, non évaluées ici, ont recours à des travailleurs humains pour compléter le travail algorithmique : les « travailleurs du clic » (Casilli, 2019). Or de plus en plus de travaux sur ces emplois documentent globalement des conditions de travail désastreuses : sur Amazon Mechanical Turk par exemple, 52 % des travailleurs

américains estiment recevoir moins de 5 $ par heure quand le salaire minimum est établi à 7,25 $ (Hitlin, 2016).

Pour finir, une dernière remarque porte sur le travail que nous offrons, nous-mêmes, à ces plateformes. De la même manière que nous travaillons pour le supermarché lorsque nous scannons nous-mêmes nos articles à une caisse automatique, les plateformes nous mettent à contribution sans le dire explicitement. En effet lorsque nous éditons notre texte en ligne, dans l'éditeur intégré à la plateforme, le couple audio/transcription obtenu sera utilisé par la plupart d'entre elles pour améliorer leurs algorithmes. Or ce travail est réalisé gratuitement, implicitement, et ce malgré le prix d'un abonnement qui peut parfois être élevé.

IX. Conclusion

Ce travail a permis de montrer qu'il n'existe pas de plateforme qui serait uniformément performante, mais plutôt des groupes de plateformes spécialisées dans des types de discours particuliers. Sonix et Vocapia 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 suggèrent au final une correction manuelle des erreurs. D'après nos observations, celle-ci peut 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.

La plupart de ces outils posent en outre 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 pas seulement en termes de recherche (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, et permettrait leur déploiement à un public à l'étranger, notamment avec un couplage à des outils de traduction automatique. Le développement d'une telle offre pourrait changer le visage du marché en constante évolution de la transcription automatique, actuellement majoritairement privé.

Références

Authôt (2016) Reconnaissance automatique de la parole au coeur de l'application Authôt. Available at: https://www.authot.com/fr/2016/09/09/systeme-reconnaissance-de-la-parole/ (accessed 3 July 2020).

Bazillon T, Estève Y and Luzzati D (2008) Manual vs assisted transcription of prepared and spontaneous speech. In: *LREC 2008*, Marrakech, Morocco, 2008. Available at: https://hal.archives-ouvertes.fr/hal-01433962.

Ben Jannet MA (2015) *Évaluation adaptative des systèmes de transcription en contexte applicatif*. These de doctorat. Université Paris-Saclay (ComUE). Available at: http://www.theses.fr/2015SACLS041.

Benveniste C-B (2000) Corpus de français parlé. In: *Corpus : Méthodologie et Applications Linguistiques*. Honoré Champion. Paris, pp. 15–25.

Benzitoun C (2014) La place de l'adjectif épithète en français : ce que nous apprennent les corpus oraux. Neveu F, Blumenthal P, Hriba L, et al. (eds) *SHS Web of Conferences* 8: 2333–2348. DOI: 10.1051/shsconf/20140801066.

Bilger M (2008) *Données orales: les enjeux de la transcription*. Perpignan: Presses universitaires de Perpignan.

Bunce T (2017) Comparing Transcriptions. In: *Not this...* Available at: https://blog.timbunce.org/2017/02/09/comparing-transcriptions/ (accessed 4 July 2020).

Bunce T (2018) A Comparison of Automatic Speech Recognition (ASR) Systems. In: *Not this...* Available at: https://blog.timbunce.org/2018/05/15/a-comparison-of-automatic-speech-recognition-asr-systems/ (accessed 3 July 2020).

Bunce T (2019) A Comparison of Automatic Speech Recognition (ASR) Systems, part 2. In: *Not this...* Available at: https://blog.timbunce.org/2019/02/11/a-comparison-of-automatic-speech-recognition-asr-systems-part-2/ (accessed 3 July 2020).

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

Casilli AA (2019) *En Attendant Les Robots: Enquête Sur Le Travail Du Clic*. La Couleur des idées. Paris XIXe: Éditions du Seuil.

Commission Européenne (2019) Guide du Bouclier de Protection des données UE-États-Unis. Available at: http://ec.europa.eu/newsroom/document.cfm?doc_id=47790.

Errattahi R, El Hannani A and Ouahmane H (2018) Automatic Speech Recognition

Errors Detection and Correction: A Review. *Procedia Computer Science* 128. 1st International Conference on Natural Language and Speech Processing: 32–37. DOI: 10.1016/j.procs.2018.03.005.

Favre B, Cheung K, Kazemian S, et al. (2013) Automatic human utility evaluation of ASR systems: Does WER really predict performance? In: *INTERSPEECH*, 2013, pp. 3463–3467.

Forsgren M (2016) La place de l'adjectif épithètes. *Encyclopédie grammaticale du français*. en ligne : encyclogram.fr. Available at: http://www.encyclogram.fr/notx/009/009_Notice.php.

Habert B (2005) Portrait de linguiste (s) à l'instrument. *Revue Texto* 10(4).

Hitlin P (2016) Turkers in this canvassing: young, well-educated and frequent users. In: *Research in the Crowdsourcing Age, a Case Study*. Pew Research Center. Available at: https://www.pewresearch.org/internet/2016/07/11/turkers-in-this-canvassing-young-well-educated-and-frequent-users/.

Kerbrat-Orecchioni C (1980) *L'énonciation: de la subjectivité dans le langage / Catherine Kerbrat-Orecchioni*. Linguistique. Paris: AColin.

Kerbrat-Orecchioni C (1990) *Les interactions verbales [Texte imprimé] / Catherine Kerbrat-Orecchioni*. [Nouv. éd. revue et augm., nouv. tirage]. Linguistique. Paris: AColin.

Lamberterie I de, Baude O, Blanche-Benveniste C, et al. (2006) *Corpus oraux. Guide des bonnes pratiques 2006*. CNRS Editions. Available at: https://halshs.archives-ouvertes.fr/halshs-00078730.

Maier V (2002) Evaluating RIL as basis of automatic speech recognition devices and the consequences of using probabilistic string edit distance as input. *Univ. of Sheffield, third year project*.

McCowan IA, Moore D, Dines J, et al. (2004) On the Use of Information Retrieval Measures for Speech Recognition Evaluation. REP_WORK. IDIAP. Available at: https://infoscience.epfl.ch/record/83156.

Miller GA (1955) Note on the bias of information estimates. *Information theory and psychology*: 95–100.

Mondada L (2000) Les effets théoriques des pratiques de transcription. *Linx. Revue des linguistes de l'université Paris X Nanterre* (42). 42. Département de Sciences du langage, Université Paris Ouest: 131–146. DOI: 10.4000/linx.902.

Morris AC, Maier V,Green PD (2004) From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In: *INTERSPEECH*, 2004.

Nanjo H, University R and Kawahara T (2005) A New ASR Evaluation Measure and Minimum Bayes-Risk Decoding for Open-domain Speech Understanding. In: *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, Philadelphia, Pennsylvania, USA, 2005, pp. 1053–1056. IEEE. DOI: 10.1109/ICASSP.2005.1415298.

Park Y, Patwardhan S, Visweswariah K, et al. (2008) An empirical analysis of word error rate and keyword error rate. In: *Ninth Annual Conference of the International Speech Communication Association*, 2008.

R Core Team (2020) *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Vienna, Austria. Available at: https://www.R-project.org/.

Rioufreyt T (2016) La transcription d'entretiens en sciences sociales. Available at: https://halshs.archives-ouvertes.fr/halshs-01339474.

Rioufreyt T (2018) La transcription outillée en SHS. Un panorama des logiciels de transcription audio/vidéo. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 139(1). SAGE Publications Ltd: 96–133. DOI: 10.1177/0759106318762455.

Traverso V (2007) *L'analyse des conversations*. 128 Lettres Linguistique. Paris: Armand Colin.

Wang Y-Y, Acero A and Chelba C (2003) Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy. Available at: https://www.microsoft.com/en-us/research/publication/is-word-error-rate-a-good-indicator-for-spoken-language-understanding-accuracy/.

Woodard J and Nelson J (1982) An information theoretic measure of speech recognition performance. In: *Workshop on standardisation for speech I/O*, 1982.

Listes des tableaux et figures

Tableaux

Figures

Annexes

Annexe 1 — Procédures de normalisation

Avant de comparer chacune des transcriptions automatiques à son pendant originel, un travail d'harmonisation a été entrepris sur chacun des fichiers utilisés dans le cadre de notre corpus. Nous présentons ici les différentes étapes d'harmonisation.

Création des fichiers REF

Harmonisation des transcriptions originales

Les transcriptions originales qui composent notre corpus proviennent de transcripteurs différents et correspondent à des situations interactionnelles hétérogènes. Avant d'opérer des traitements dessus, nous souhaitions vérifier la fidélité de la transcription par rapport aux extraits audio correspondants.
Chacun des quatre enregistrements a ainsi été réécouté par un transcripteur expert, qui a vérifié les quatre transcriptions originelles dans le but de les harmoniser.
Par ailleurs, nous avons appliqué une convention de transcription commune à notre corpus. En particulier, le choix de transcrire en orthographe dite « standard » a été retenu.
Ce travail a été effectué à l'aide de logiciels comme TextWrangler (version Mac) qui nous a permis d'éditer et de corriger la transcription. Le logiciel Audacity quant à lui, nous a permis d'écouter le signal audio.

Les corrections ont porté sur un certain nombre d'opérations que nous détaillons ci-après. Certaines étaient communes à l'ensemble du corpus, d'autres sont spécifiques à certains sous-corpus.

Traitements communs

Les traitements communs concernaient la vérification lexicale et l'harmonisation des conventions de transcription utilisées pour les sous-corpus. Un double objectif était visé : tenter d'obtenir des fichiers uniformes en matière de convention de transcription et garantir autant que possible la fidélité de restitution par rapport aux supports originaux. Ces vérifications ont porté en particulier sur la concordance du lexique par rapport au fichier audio, la conservation des phénomènes linguistiques comme les répétitions (*il il a décidé*), les reformulations ainsi que les amorces/troncations (*mercre- mercredi*), la conservation des interjections comme *hein*, *ben*, *hum*. Concernant les *euh* présents en grand nombre dans certains extraits, nous avons pris la décision de supprimer la majeure partie d'entre eux afin de faciliter l'analyse, tout en conservant quelques occurrences saillantes dans le but d'étudier la capacité des logiciels à les traiter. Nous avons également supprimé les annotations diverses (prosodie, pauses…) lorsqu'elles étaient présentes. Enfin nous avons procédé à une harmonisation typographique des guillemets et des apostrophes (remplacement des « courbes » par des « droits ») et procéder à la suppression des espaces surnuméraires en particulier dans les élisions).

Les fichiers ont ensuite été enregistrés au format texte Windows (adoption des passages à la ligne type CR & LF) et encodés en UTF-8. Une copie au format RTF a été conservée afin de s'affranchir des problèmes d'encodage rencontrés dans certains logiciels (par exemple avec la version Macintosh de TexEdit).

La nature des extraits étant très diverse, certains d'entre eux ont nécessité un traitement supplémentaire.

Camille

– segmentation des lignes selon les tours de parole, à raison d'un saut de ligne pour chaque changement de locuteur (Laure/Camille) ou lorsqu'un chevauchement se produisait
– chaque tour de parole débute par une majuscule (sauf pour les chevauchements de façon à les distinguer et conserver la continuité du discours),
– conservation de la ponctuation en supprimant celle qui paraissait ambiguë ainsi que les points de suspension. Ces derniers étaient essentiellement utilisés pour noter les allongements,
– notation des heures sous la forme : *9 heures 30* ; les adresses ; les années : *2020*,
– rectification de quelques coquilles de transcription en privilégiant les mots/propositions tels qu'ils ont été prononcés (ex : *que on* plutôt *qu'on*),
– conservation des amorces (une seule occurrence dans l'extrait : *mercre* pour *mercredi*).
– suppression de la plupart des occurrences de *euh* dans l'extrait

Comptines

– réduction des passages à la ligne pour éviter un texte trop fragmenté. Cette segmentation s'est basée principalement sur la prosodie, les pauses significatives ou les transitions thématiques.

Harmonie

– Cet extrait se caractérise par une présence importante de chevauchements et de passages inaudibles. Ces derniers ont été notés par la mention (inaud.),
— beaucoup de *euh* (une petite trentaine). *Nous avons conservé quelques occurrences jugées saillantes placées en début de fichier.*
– il contient un passage anonymisé sous forme de BEEP, celui-ci a été noté sur la transcription par *BEEP*

Harmonisation formelle

Une autre personne du groupe a ensuite harmonisé formellement ces nouveaux fichiers, à l'aide du logiciel TextWrangler. Les fins de lignes ont été normalisées, les caractères non ASCII, caractères de contrôles et caractères NULL ont été identifiés et remplacés, afin de nous assurer qu'ils étaient encodés de manière identique dans chacun des fichiers. Les espaces ont été harmonisées selon les règles typographiques du français de France. Les ligatures (œ, Œ) ont été rétablies quand elles manquaient.

Enfin, les espaces superflues ont été supprimées. Les fichiers ont ensuite été enregistrés en UTF-8, avec des sauts de lignes Windows (CRLF).

Harmonisation des fichiers RES

En complément du travail effectué sur les fichiers **REF**, chaque membre du projet a hormonisé les fichiers **RES** qu'il avait obtenus lors de ses transcriptions automatiques, en suivant les consignes discutées collectivement :

- enlever noms de locuteurs et balises temporelles en sortie des outils ;
- conserver les *euh* et les répétitions ;
- harmoniser les *heu* (*heu*, *euh*) ;
- conserver la ponctuation ;
- garder les chiffres ou les chiffres écrits en toutes lettres ;
- supprimer les espaces après les apostrophes et les espaces surnuméraires ;
- changement des apostrophes courbes en apostrophes droites ;
- enregistrer en .TXT, UTF8, sauts de ligne Windows.

Annexe 2 — Script R

```
# script de mesure de distance entre textes
# projet "La transcription automatique : un rêve enfin accessible ?"
# Elise Tancoigne, 7 avril 2020

# chargement des packages nécessaires
require(RCurl) # pour récupérer les données sur le serveur
require(tidyr) # pour utiliser les pipes %>%
require(tm) # pour les fonctions de nettoyage de texte
require(quanteda) # pour le calcul du WER
require(stringr) # pour le calcul du WER

# RECUPERATION DES DONNEES SUR UN SERVEUR DISTANT

URL_UTBOX1 <- "url_dossier_textes_reference"
URL_UTBOX2 <- "url_dossier_transcriptions_originales_harmonisees"
URL_UTBOX3 <- "url_dossier_transcriptions_plateformes_harmonisees"


# liste des plateformes et des fichiers
platform  <-  c("goTRANSCRIBE",  "HAPPYSCRIBE",  "SONIX",  "VI-MICROSOFT",
"YOUTUBE", "HEADLINER", "VOCAPIA", "VOCALMATIC")
file    <-    c("5MIN_Camille",    "5MIN_Physionomie",    "5MIN_Harmonie",
"5MIN_Comptines")

# on crée un tableau vide fichier x plateformes
data = data.frame(matrix(vector(), 0, 10,
                         dimnames=list(c(),    c("reference",    "manuelle",
platform))),
                  stringsAsFactors=F)

# on remplit le tableau avec les textes téléchargés du serveur

# dans la première colonne : fichier de transcriptions de références
# pour chaque fichier :
for (j in 1:length(file)){
  # on concatène les URLs et on récupère le fichier
  data[j,1] <- getURL(paste(URL_UTBOX1, file[j], "_revu.txt", sep=""))
}
row.names(data) <- file

#on visualise les données
View(data)

#dans la deuxième colonne : fichiers de transcriptions originales (harmonisées
pour le WER)
#pour chaque fichier :
for (j in 1:length(file)){
  # on concatène les URLs et on récupère le fichier
  data[j,2] <- getURL(paste(URL_UTBOX2, file[j], "_WER.txt", sep=""))
}

# puis pour chaque plate-forme, on récupère le fichier transcrit (et harmonisé)
correspondant
for (i in 1:length(platform)){
  # pause d'1 seconde pour ne pas charger le serveur
  Sys.sleep(1)
  # affichage de la progression par plateforme
  print(paste(c("Processing", platform[i], sep=" ")))
  # pour chaque fichier
  for (j in 1:length(file)){
    # affichage de la progression par fichier
    print(paste(c("Processing", file[j], sep=" ")))
    # on concatène les URLs pour récupérer le fichier
    data[j,i+2] <- getURL(paste(URL_UTBOX3, platform[i], "&files=", file[j],
"_", platform[i], "_brut_harmonise.txt", sep=""))
```

```
    # attention il faut mettre i+2 pour ne pas écraser les deux premières
colonnes (transcriptions de référence et manuelles)
  }
}
rm(i,j)

#on visualise les données
View(data)

# NETTOYAGE DES DONNEES

# on copie le dataset pour travailler dessus et garder l'original
dataCleaned <- data

# on cree deux fonctions
# -> une fonction pour nettoyer les fichiers originaux
cleanDataOrig = function(text){
  # on supprime les retours ligne
  text <- gsub('\r|\n', " ", text)
  # on met tout en minuscule, on supprime la ponctuation, les nombres, les
espaces surnuméraires
  text %>%
    tolower() %>%
    removePunctuation() %>%
    removeNumbers() %>%
    stripWhitespace() %>%
    return()
}
# -> une fonction pour nettoyer les fichiers transcrits
cleanData = function(text){
  # on supprime le titre si besoin
  if (grepl("harmonisé|Aegisub", text)){
    text <- gsub('.+harmonis.{1,5}\n|# Exported by Aegisub 3.2.2', " ", text)
  }
  # on met tout en minuscule, on supprime la ponctuation, les nombres, les
espaces surnuméraires
  text %>%
    tolower() %>%
    removePunctuation() %>%
    removeNumbers() %>%
    stripWhitespace() %>%
    return()
}

# on réalise le nettoyage

# pour les fichiers de référence et les transcriptions manuelles
for (j in 1:4){
  dataCleaned[j,1] <- cleanDataOrig(dataCleaned[j,1])
  dataCleaned[j,2] <- cleanData(dataCleaned[j,2])
}
dataCleaned[3,1] <- gsub("inaud ", "", dataCleaned[3,1])

# pour les transcriptions issues des plateformes
# pour chaque plateforme
for (i in 3:10){
  # pour chaque fichier
  for (j in 1:4){
    print(c(i, j))
    dataCleaned[j,i] <- cleanData(dataCleaned[j,i])
  }
}

View(dataCleaned)


# CALCUL DU WER
```

```
# fonction développée par Jens Wäckerle, package wersim
# disponible sur https://rdrr.io/github/jenswaeckerle/wersim/src/R/wer.R
# mise à jour pour la rendre fonctionelle

wer<-function(r,h){
  ##### Errors
  # Add an error if corpora have different lengths
  if (length(r)!=length(h))
    stop("The refernce and hypothesis corpus should have the same length")

  if(length(r)==length(h)){
    data.store=data.frame(wer=rep(NA,length(r)),
                          sub=NA,ins=NA,del=NA,words.ref=NA,words.hyp=NA)
    for(k in 1:length(r)){
      print(paste("Document",k,"of",length(r)))
      sub.count=0
      ins.count=0
      del.count=0
      ref_text=tolower(unlist(stringr::str_split(r[k]," ")))
      hyp_text=tolower(unlist(stringr::str_split(h[k]," ")))
      if(h[k]==""){
        data.store$wer[k]=1
        data.store$sub[k]=0
        data.store$del[k]=length(ref_text)
        data.store$ins[k]=0
        data.store$words.ref[k]=length(ref_text)
        data.store$words.hyp[k]=0
      }
      if(h[k]!=""){
        d1<-matrix(ncol=length(hyp_text)+1,nrow=length(ref_text)+1,0)
        d1[1,]<-0:length(hyp_text)
        d1[,1]<-0:length(ref_text)
        dtext=d1
        for(i in 2:nrow(d1)){
          for(j in 2:ncol(d1)){
            if(ref_text[i-1]==hyp_text[j-1]){
              d1[i,j]<-d1[i-1,j-1]
              dtext[i,j]="CORRECT"
            }
            else{
              sub<-d1[i-1,j-1]+1
              ins<-d1[i,j-1]+1
              del<-d1[i-1,j]+1
              d1[i,j]<-min(sub,ins,del)
              if(which.min(c(sub,ins,del))==1){
                dtext[i,j]="SUB"
              }
              if(which.min(c(sub,ins,del))==2){
                dtext[i,j]="INS"
              }
              if(which.min(c(sub,ins,del))==3){
                dtext[i,j]="DEL"
              }
            }
          }
        }
        sequence=rep(NA,length(ref_text))
        start.row=nrow(dtext)
        start.col=ncol(dtext)
        dtext[2:nrow(dtext),1]="DEL"
        dtext[1,2:ncol(dtext)]="INS"
        dtext[1,1]="CORRECT"
        for(l in (length(sequence)):1){
          sequence[l]=dtext[start.row,start.col]
          if(sequence[l]%in%c("CORRECT","SUB")){
            start.row=start.row-1
            start.col=start.col-1
          }
```

```
          if(sequence[l]=="DEL"){
            start.row=start.row-1
          }
          if(sequence[l]=="INS"){
            start.col=start.col-1
          }
        }

data.store$wer[k]=d1[length(ref_text)+1,length(hyp_text)+1]/length(ref_text)
        data.store$sub[k]=sum(sequence=="SUB",na.rm=T)
        data.store$del[k]=sum(sequence=="DEL",na.rm=T)
        data.store$ins[k]=sum(sequence=="INS",na.rm=T)
        data.store$words.ref[k]=length(ref_text)
        data.store$words.hyp[k]=length(hyp_text)
      }
    }
    return(data.store)
  }
}

# on crée un tableau vide fichier x plateformes
dataWER <- data.frame(matrix(vector(), 4, 9,
                             dimnames=list(c(), c("manuelle", platform))),
                      stringsAsFactors=F)

# on calcule la mesure pour la retranscription manuelle
dataWER[,1] <- wer(corpus(dataCleaned[,2]), corpus(dataCleaned[,1]))$wer

# on calcule la mesure pour les 8 plateformes
for (i in 1:length(platform)){
  #on affiche laquelle on traite
  print(platform[i])
  # on compare la plateforme avec le texte d'origine
  temp <- wer(corpus(dataCleaned[,i+2]), corpus(dataCleaned[,1]))
  # on remplit la colonne correspondante
  dataWER[,i+1] <- temp$wer
}
rm(temp, i)
row.names(dataWER) <- file
View(dataWER)
# on enregistre le fichier
write.csv(dataWER, "chemin_dossier/dataWER.csv")

# on enregistre l'environnement
save.image("chemin_dossier/WER&Co.RData")
```

Annexe 3 — Guide d'annotation

Sommaire 1

Remarque : la comparaison réalisée porte uniquement sur le lexique, la ponctuation est laissée de côté, tout comme la casse. Toutes deux ne sont par ailleurs pas comptabilisées par Copyscape[34].

Remarque : un champ remarque est disponible dans les fichiers Excel. Il sert à relever des exemples remarquables, à préciser ses choix en cas d'hésitation et à justifier l'utilisation de la catégorie « autres ».

Remarque : nous appellerons **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 1 mot).

Problème : Comment compter *9 h 30* ? Un mot ou trois ? On comptera ici trois, car nous avons privilégié la convention de notation *9 heures 30* dans les transcriptions de référence.

Mots de RES identiques à REF

Certains passages des fichiers RES sont considérés par Copyscape comme étant identiques à ceux du fichier REF. L'outil fournit alors un nombre de mots pour ces passages. Ce nombre est reporté tel quel, après vérification du décompte, comme dans la Figure 1 ci-dessous[35]. Notons que dans cette figure, un comptage manuel nous aurait amenés à comptabiliser 22 mots communs et non pas 23 comme l'a fait Copyscape.

[34] Sauf cas particuliers, comme un cas de « ; » observé dans Physionomie.
[35] Remarque : dans les tableaux qui suivent la première colonne correspond au fichier REF, la troisième au fichier RES.

701 words, 68% matched		702 words, 68% matched	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	mots de REF absents de RES	mots de RES qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	erreurs de flexion	autre
De la Physionomie.		de la physionomie		0		0	0			0	0
CHAPITRE 12		chapitre 12		0		0	0			0	0
Quasi toutes les opinions que nous avons sont prises par autorité et à crédit. Il n'y a		quasi toutes les opinions que nous avons sont prises par autorité et à crédit		0		0	0			0	0
	« 23 words »	Il n'y a	23	0		0	0			0	0

Figure 1 Passages identiques dans fichier RES et REF selon Copyscape

Seuls les passages considérés par Copyscape comme étant distincts font l'objet d'un décompte et d'une classification. Certains d'entre eux comportent des mots identiques dans les deux fichiers comparés. C'est le cas du mot *qui* dans le passage illustré par la Figure 3, ci-dessous.

701 words, 68% matched		702 words, 68% matched	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	mots de REF absents de RES	mots de RES qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	erreurs de flexion	autre
préceptes, qui		précepte qui est	1	0		0	1			1	0

Figure 3 « préceptes, qui » — transcription Headliner

Seuls les mots strictement identiques sont comptabilisés ici. Les différences de conventions de notation des chiffres, des mois ou des unités de mesure ne sont pas prises en compte. Ainsi *2* et *deux* sont considérés comme strictement identiques, tout comme *9 heures 30* et *9 h 30*.

Erreurs de flexion

Lemme : forme canonique d'un mot variable (par convention : masculin singulier d'un adjectif, infinitif d'un verbe, singulier d'un nom commun)
Flexion : forme d'un mot variable telle qu'on la trouve dans un texte (avec des marques de genre, de nombre, de temps, etc.)

Certains mots diffèrent d'un fichier REF à un fichier RES uniquement en termes de flexions (changement de nombre, de genre, de conjugaison), comme c'est le cas de *au*, *épineuse*, *traverse* et *puisse* dans la Figure 3 ci-dessous.

Item 1		Item 2	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	mots de REF absents de RES	mots de RES qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	erreurs de flexion	autre
aux plus épineuses traverses qui se puissent		au plus épineuse traverse qui se puisse	3	0		0	0			4	0

Figure 34 « aux plus épineuses traverses qui se puissent » — transcription Happy Scribe

Dans ce cas de figure, nous ne considérons pas que nous sommes en présence de mots identiques, parce que la transcription produite est incorrecte et qu'une correction va devoir être apportée.

Ces erreurs peuvent induire des différences de sens importantes (*celles qui coulent échappent* vs *celle qui coule échappe*), mais celles-ci nuisent moins à la compréhension générale du discours que lorsque le lemme lui-même est incorrect (cf. mauvais lexique).

Attention, on parle bien ici d'erreurs de flexion et non de familles morphologiques. Ainsi, dans le cas de *se monta* transcrit *ce montant*, nous ne compterons aucune erreur de flexion : *monta* est un verbe et à pour lemme *monter*, tandis que *montant* est ici un nom commun est à pour lemme *montant*.

Il est parfois difficile de déterminer si le lemme est bien le même dans les fichiers REF et RES. Dans le passage illustré par la Figure 3, il est clair que *traverses* est un nom dans l'item 1. Il est en revanche difficile de déterminer si *traverse* est employé comme verbe (*traverser*) ou comme nom (*traverse*) dans l'item 2. On privilégie alors la classification comme erreurs de flexion.

Exemples

laissée → laissé
bouffies → bouffis
s'enflent → s'enfle
manient → manie
dressa → dresse
monta → monte
boutades → boutade
il → ils

Substitution avec proximité phonétique

Certains mots apparaissent dans un fichier RES, mais sont absents du fichier REF. Il est cependant clair qu'il ne s'agit pas de mots ajoutés, mais d'erreurs de choix d'unités lexicales lors de la transcription. Nous les catégorisons comme des substitutions. Nous comptons ici le nombre de mots produits dans le fichier RES, quelle que soit la quantité de mots correspondant dans le fichier REF. Nous distinguons toutefois les cas où l'on perçoit une proximité phonétique entre la transcription de référence et la proposition de transcription faite par la transcription automatique, comme dans le passage illustré par la Figure et ceux où l'on n'en perçoit pas. Ici les mots *m'ont*, *pas* et *Ravana* sont comptés comme 3 cas de substitution avec proximité phonétique.

Item 1		Item 2	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	mots de REF absents de RES	mots de RES qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	erreurs de flexion	autre
il ne monta rien, mais ravala		Ils ne m'ont pas rien, mais Ravana	3	3		0	0			1	0

Figure 4 « il ne monta rien, mais ravala » — transcription Sonix

Exemples

sont → *son*
vaines → *veines*
cochers → *kosher*
l'homme → *Lomme*
aux braves → *Aubrac*
cette heure → *7 h* (1 substitution + 1 identique)
33 → *endroit*
et pile à → *épinal*
à ces → *assez*
à notre → *un autre*
en dehors → *endort*
dénombrer → *des nombres et* (3 substitutions)
leur jeu sans → *l'heure je sens* (3 substitutions)
moyenne on met → *moyen non, mais* (3 substitutions)
par ces → *parce et* (2 substitutions)
elle perdait son temps → *est le père des sont en* (6 substitutions)
et sans se → *essence* (1 seule substitution)
euh en dehors de l'accompagnement administratif → *à Andorre de la compagnie Mani stratif* (1 mot identique, 6 substitutions)

Substitution sans proximité phonétique

En l'absence de proximité phonétique, on s'interroge sur la possibilité d'aligner les mots présents dans les fichiers RES avec ceux du fichier REF. Dans le cas de la Figure , on peut estimer que *recevant* correspond à la fin de *n'apercevant* et que cette substitution relève d'un cas avec proximité phonétique. On n'a alors pas grande difficulté à accepter que *fait* a été mis pour le même segment de signal que le début de *n'apercevons*. Un tel cas relève de la catégorie substitution sans proximité phonétique.

699 words, 66% matched		701 words, 66% matched	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	mots de REF absents de RES	qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	erreurs de flexion	autre
Nous n'apercevons		nous fait recevant	1	1	1	0	0			0	0

Figure 5 « nous n'apercevons » — transcription Vocalmatic

Exemples

dans 20 mn → *très vite* (2 substitutions)
Excuses → *ce*
Monsieur → *c'est*
mairie → *vallée*
voilà → *fort*

Passage sans lien (nombre de mots REF/nombre de mots RES)

Dans certains cas, en revanche, il est difficile, voire impossible de réaliser un alignement entre les mots du fichier REF et ceux produits dans le fichier RES. En présence de tels passages, nous adoptons la position suivante :

- compter le nombre de mots RES identiques à ceux de REF
- noter en remarque « aucun lien perçu entre REF et RES »
- compter les autres mots de REF comme nbr de mots REF
- compter les autres mots de RES comme nbr de mots REF

Ainsi, dans la Figure , *est pile à cheval* et *je vais le faire*, sont comptés comme 4 mots REF et 4 mots RES, tandis que *sur les deux communes* et *sur les 2 communes* sont comptabilisés comme mots de RES identiques à REF.

661 words, 72% matched		619 words, 77% matched	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	mots de REF absents de RES	mots de RES qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	erreurs de flexion	autre
est pile à cheval sur les deux communes		je vais le faire sur les 2 communes.	4	0		0	0	4	4	0	0

Figure 6 « est pile à cheval sur les deux communes » — transcription Vocapia

Dans certains cas, on souhaitera également identifier un ou plusieurs mots comme étant phonétiquement proches. On utilisera alors la catégorie substitution avec proximité phonétique, en plus des deux catégories de passage sans lien. Dans la Figure , on peut voir un tel cas de figure, avec un alignement supposé de *Cassandra* et de *ça sonnera*.

			mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	erreurs de flexion	mots de REF absents de RES	mots de RES qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	autre
ça sonnera dans 20 mn.		Cassandra dans 20 minutes	3	1		0	0			0	0

Figure 7 « ça sonnera dans 20 mn. » — transcription Vocalmatic

Mots de REF absents de RES

Certains mots présents dans le fichier REF ne sont clairement pas retranscrits dans le fichier RES. C'est le cas du mot *saine* dans la Figure ci-dessous.

701 words, 79% matched		700 words, 79% matched	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	mots de REF absents de RES	mots de RES qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	erreurs de flexion	autre
ni riche : il ne la représente que	« 7 words »	ni riche.		0		0	0			0	0
		Il ne la représente que	7	0		0	0			0	0
saine,			0	0		1	0			0	0
mais certes d'une bien allègre et	« 6 words »	mais certes d'une bien Allègre et	6	0		0	0			0	0

Figure 8 « saine » — transcription Vocapia

Dans certains cas, il peut être difficile de déterminer si on est en présence de mots absents ou non, comme dans dans *essence, piqué*[36] transcrit à la place de *et sans se piquer*. Tant qu'un alignement nous semble possible et justifiable, nous privilégierons la substitution. Ici nous faisons donc l'hypothèse que *et sans se* a été transcrit *essence* et nous comptons pour cette correspondance 1 substitution avec proximité phonétique et 1 erreur de flexion (*piqué* pour *piquer*).

La majorité des plateformes étudiées ne transcrivent pas les onomatopées et interjections. Seules les occurrences de *hein* et de *hum* doivent être systématiquement comptées comme absentes si elles le sont. Se référer à la section Onomatopées et interjections pour davantage de précisions.

Mots de RES qui ne correspondent à rien dans REF (ajout)

Certains mots présents dans les fichiers RES ne sont clairement pas présents dans le fichier REF correspondant. C'est le cas du mot *le* dans la Figure 9 ci-dessous.

Item 1 701 words, 76% matched		Item 2 694 words, 77% matched	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	mots de REF absents de RES	qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	erreurs de flexion	autre
cachée: il faut la vue nette et bien purgée		caché. Il faut l'as vu le net et bien purger	4	3		0	1			2	0

Figure 9 « la vue nette » — transcription Video indexer

Dans certains cas, il est difficile de déterminer si on est en présence d'un ajout ou non, comme dans *Lomme où est l'état ça pue juste* transcrit à la place de *l'homme, où est sa plus juste*. On privilégiera alors la substitution.

Exemples

(le mot compté comme ajouté est en gras)

*d'emprunté → **de** emprunter*
*n'apercevons → **ne** percevons*
*sont inductions → sont **à** induction*

Autres

La catégorie autres sert à classer les cas particuliers que l'on n'a pas envie de classer dans une autre catégorie. Son usage doit être accompagné d'un commentaire.

À l'issue de la classification, nous retenons trois sous-catégories d'éléments répertoriés dans la catégorie « autres » : des onomatopées et interjections présentes dans les fichiers REF et absentes des fichiers RES ; des cas d'élision ; des cas divers.

[36] exemple tiré de la transcription de Physionomie par VI-Microsoft.

Onomatopées et interjections

Nous comptabilisons dans la catégorie autres les onomatopées *euh* présentes dans le fichier RES ou le fichier REF, mais absentes du second[37]. Ces onomatopées marquent l'hésitation et la temporisation dans le discours, mais elles n'ont généralement pas de sens défini, comme le montre la Figure 5.

661 words, 66% matched		609 words, 71% matched	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	mots de REF absents de RES	mots de RES qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	erreurs de flexion	autre
Euh non pas Camille		non pas qu amy	2	2		0	0			0	1

Figure 5 euh non pas Camille — transcription YouTube

Selon la perspective de l'analyse interactionnelle, les *hein* exercent une fonction phatique dans le discours (Kerbrat-Orecchioni, 1980, 1990 ; Traverso, 2007) et selon le contexte du discours peuvent se substituer à d'autres mots comme *quoi*. C'est pourquoi nous en comptabilisons les occurrences non transcrites comme des « mots absents », comme dans la Figure .

Item 1 *837 words, 86% matched*		*Item 2* *810 words, 89% matched*	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	erreurs de flexion	mots de REF absents de RES	mots de RES qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	autre
la variété des projections qu'il peut y avoir sur les poupées russes	12	la variété des projections qu'il peut y avoir sur les poupées russes.	12								
hein							1				
chacun y met ses significations											

Figure 11 hein — Go Transcribe

Le cas des *bah* et *ben* est plus délicat et leurs occurrences sont traitées de façon moins systématique. Les choix suivants ont été effectués :

- occurrence de *bah* dans Comptines : mot de REF absent de RES, car dans le contexte de la phrase il a le sens de *bien*. On peut même s'aventurer à dire que c'est une contraction, une altération de *bien ;*
- occurrences de *ben* dans Camille : toutes deux mots de REF absent de RES[38], car on considère ici qu'elles correspondent à *et bien* ;
- occurrences de *ben* dans Harmonie :
    - la première (*ben ça je le laisse avec ça*) comptée comme autres car il est quasiment impossible de certifier que c'est un *ben*, un *bah* ou autre chose qui est prononcé dans le signal.
    - la seconde (*s'il y a pas de démissionnaire et ben on revote*) comptée comme mot de REF absent de RES, car à l'écoute du signal, c'est clairement un *ben* qui substitue un *bien*

[37] Ces onomatopées sont majoritairement présentes dans le fichier REF et absentes des fichiers RES, mais l'inverse a également été observé, notamment dans les transcriptions proposées par VI-Microsoft.
[38] La catégorie peut être un sous-type de substitution si la plateforme a proposé une transcription, comme c'est le cas pour YouTube qui propose *ai ensuite berne* pour le segment *et ensuite ben*.

- la troisième (*ben on commence par le*) comptée comme autres, car difficilement audible dans le l'extrait audio.

Élision

Élision : suppression de la voyelle finale d'un mot qui précède un mot commençant par une voyelle ou un *h* muet.

Remarque : le nombre de autres comptabilisés correspond au nombre de mots du fichier RES selon les règles établies précédemment : *s'est* compte pour 1, *d'une autre* pour 2.

Pour coller au mieux au décompte de mots communs de Copyscape, nous avons décidé que l'apostrophe ne constituait pas un délimiteur de mot. Dans certains cas, en raison du phénomène d'élision fréquent en français, ce choix nous empêche de classer certaines erreurs comme on le souhaiterait. Ainsi, on peut voir dans la Figure , un cas où *l'enfant* a été transcrit *d'enfants*. Sans élision, nous aurions deux mots dans le fichier REF (*le*, *enfant*) et dans le fichier RES (*de*, *enfants*) et nous pourrions compter un cas de substitution et un cas d'erreur de flexion. Faute de pouvoir réaliser ce classement, nous avons opté pour une catégorisation en un cas autre.

Item 1	Nombre d	Item 2	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	erreurs de flexion	mots de REF absents de RES	mots de RES qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	autre
837 words, 70% matched		**736 words, 79% matched**									
l'enfant		d'enfants									1

Figure 6 l'enfant — transcription Headliner

La présence d'élision ne déclenche cependant pas systématiquement un classement des erreurs en autre. Dans le cas présenté dans la Figure 7Figure 7, nous aurions, sans élision, deux mots dans le fichier RES (*que*, *ils*), mais aucun ne serait identique au mot correspondant dans le fichier REF (*qui*). Nous ne comptons alors aucun autre, mais plutôt un cas de substitution.

661 words, 66% matched		609 words, 71% matched	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	mots de REF absents de RES	mots de RES qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	erreurs de flexion	autre
qui viennent à ces permanences ?		qu'ils viennent ces permanences heures	3	1		1	1			0	0

Figure 7 qui viennent à ces permanences ? — transcription YouTube

Dans le cas présenté dans la Figure 8, *t'appelle* est classé comme erreur de flexion et nous ne comptons qu'un seul autre, celui correspondant à l'onomatopée *euh* non transcrite.

			mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	mots de REF absents de RES	mots de RES qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	erreurs de flexion	autre
Camille, d'accord je t'appellerai Camille. Euh Camille		Cami D'accord, je t'appelle Cami.	2	2		1	0			1	1

Figure 8 Camille, d'accord je t'appelerai Camille. Euh Camille — transcription Sonix

Enfin, dans la Figure , nous classons l'erreur comme un cas de substitution, parce qu'il ne s'agit pas là d'un phénomène d'élision. En effet, en synchronie, *d'accord* forme une seule unité lexicale et, en cas de lemmatisation, nous conserverions *d'accord* et non *de accord*.

661 words, 57% matched		487 words, 77% matched	mots de RES identiques à REF	substitution avec proximité phonétique	substitution sans proximité phonétique	mots de REF absents de RES	mots de RES qui ne correspondent à rien dans REF (ajout)	nbr mots REF	nbr mots RES	erreurs de flexion	autre
d'accord		D'Accor	0	1		0	0			0	0

Figure 15 d'accord — transcription Go Transcribe

Exemples

c'est → *s'est*
aisément → *t'aisément*
n'eussions → *n'étions*
d'accessibilité → *de l'accessibilité* (2 autres)
t'es → *tu est* (2 autres)
de notre → *d'une autre* (2 autres)

Divers

D'autres phénomènes ponctuels ont été classés dans la catégorie « autre ». Ils sont listés ici :

point → . [pris comme une instruction d'inscrire le signe typographique]

y a → *ya* [agglutination accidentelle ou mauvais lexique ?]

Références

Kerbrat-Orecchioni C (1980) *L'énonciation : de la subjectivité dans le langage / Catherine Kerbrat-Orecchioni*. Linguistique. Paris : AColin.

Kerbrat-Orecchioni C (1990) *Les interactions verbales [Texte imprimé] / Catherine Kerbrat-Orecchioni*. [Nouv. éd. revue et augm., nouv. tirage]. Linguistique. Paris : AColin.

Traverso V (2007) *L'analyse des conversations*. 128 Lettres Linguistique. Paris : Armand Colin.

Septembre 2020