

Ni doublage, ni voice-over [br]Le doublage automatique de YouTube à l'épreuve du gaming et de la vulgarisation scientifique (anglais-français)

Auteur : Meekers, Samuel

Promoteur(s) : Wong García, Ernesto

Faculté : Faculté de Philosophie et Lettres

Diplôme : Master en traduction, à finalité spécialisée

Année académique : 2025-2026

URI/URL : <http://hdl.handle.net/2268.2/26870>

Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.

Université de Liège

Faculté de Philosophie et Lettres

Département de Langues modernes : littérature, linguistique, traduction

Ni doublage, ni voice-over

*Le doublage automatique de YouTube à
l'épreuve du gaming et de la vulgarisation
scientifique (anglais-français)*

Mémoire présenté par Samuel MEEKERS

en vue de l'obtention du grade de Master en Traduction, à finalité spécialisée

Promoteur·rice : Ernesto WONG GARCÍA, Anne DEBRAS

Année académique 2025/2026



Déclaration d'authenticité

Je soussigné·e, Samuel MEEKERS déclare avoir rédigé le présent travail de fin d'études de manière autonome, sans l'aide non autorisée de tiers, et ne pas avoir utilisé d'autres moyens que ceux indiqués. J'ai indiqué la référence précise de tout passage emprunté textuellement à d'autres sources ou paraphrasé à partir de celles-ci, peu importe leur forme ou support (p.ex. ouvrages, podcasts, vidéos, sites web).

Je déclare avoir pris connaissance de la Charte ULiège relative à l'utilisation des intelligences artificielles génératives (IA génératives) dans les travaux universitaires

(https://www.student.uliege.be/cms/c_19230399/fr/faq-student-charte-uliege-d-utilisation-des-intelligences-artificielles-generatives-dans-les-travaux-universitaires). Je comprends que par « IA génératives », on entend les agents conversationnels de type ChatGPT ainsi que tout autre outil d'IA permettant la génération ou la synthèse de texte, ou l'interprétation de données, y compris les réponses fournies automatiquement par des moteurs de recherche.

Je déclare que mon travail implique (cochez la case appropriée) :

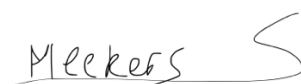
☐ Aucun usage de l'IA générative.

☒ Un usage de l'IA générative comme aide à la correction syntaxique / morphologique / lexicale.

Je suis conscient·e du fait que toute utilisation d'IA générative pour la rédaction du TFE qui dépasse celle mentionnée ci-dessus est interdite. Ceci comprend, par exemple, l'usage d'IA comme générateur des passages complets ou comme aide à la structuration ou reformulation de passages que j'ai rédigés. Je comprends que le cadre fixé par cette déclaration – qui stipule les restrictions propres à ma filière d'étude – est plus stricte que celui fixé par la Charte institutionnelle.

Ce travail peut être vérifié pour le plagiat et l'utilisation des intelligences artificielles génératives à l'aide du logiciel approprié. Je comprends qu'une conduite contraire à l'éthique peut entraîner une sanction.

Liège, le 11 août 2026,



Remerciements

Je tiens à remercier mon promoteur, Monsieur Ernesto Wong García, ainsi que ma copromotrice, Madame Anne Debras, pour leur disponibilité tout au long de ce travail et pour la justesse de leurs conseils, sans lesquels ce mémoire n'aurait pas pu être mené à son terme.

Table des matières

Remerciements.....	2
Table des matières.....	2
Liste des abréviations et notations	6
INTRODUCTION GÉNÉRALE	7
Contexte	7
État de la recherche.....	8
Problématique	9
Hypothèses.....	9
Corpus.....	10
Démarche.....	11
Structure du travail.....	11
PARTIE I – CADRE THÉORIQUE	12
1. Le doublage comme illusion d'authenticité.....	13
2. La voix comme vecteur de réussite du doublage.....	16
2.1 La corporalité de la voix	17
2.2 Prosodie et synchronisation	19
3. Le doublage automatique généré par YouTube.....	21
4. Traduction automatique neuronale et oralité préfabriquée	23
4.1 De la traduction automatique statistique à l'architecture transformer.....	23
4.2 Gemini et la chaîne de traitement du doublage automatique YouTube	24
4.3 Les limites de la TAN face au discours oral et à l'oralité préfabriquée.....	25

5. Les genres analysés et leurs enjeux pour le doublage automatique.....	28
5.1 L'oralité préfabriquée : un cadre d'analyse affiné	28
5.2 Le gaming : une oralité préfabriquée à forte dimension interactionnelle	29
5.3 La vulgarisation scientifique : entre rigueur terminologique et stratégies d'engagement.....	30
5.4 Complémentarité des deux genres pour l'analyse du doublage automatique	31
PARTIE II – MÉTHODOLOGIE	32
6. Constitution du corpus	33
6.1 Critères de sélection des vidéos	33
6.2 Présentation des vidéos.....	33
6.2.1 Corpus gaming — Captain Collins (tutoriel).....	33
6.2.2 Corpus gaming — Maya (stream commenté).....	34
6.2.3 Corpus gaming — Doael (corpus complémentaire ciblé).....	34
6.2.4 Corpus vulgarisation scientifique — Big Think (monologue)	34
6.2.5 Corpus vulgarisation scientifique — World Science Festival (entretien).....	34
6.3 Sélection des extraits	35
6.3.1 Critères de sélection.....	35
6.3.2 Procédure de repérage.....	35
6.3.3 Limites méthodologiques.....	36
7. Cadre analytique et protocole d'analyse	37
7.1 Grille d'analyse : cinq catégories de critères	37
7.2 Niveau d'analyse et indicateurs observables	39
7.3 Protocole d'analyse.....	40
7.4 Convention de transcription.....	40
PARTIE III – ANALYSE DU CORPUS.....	42
8. Analyse du sous-corpus gaming	43
8.1 Présentation des extraits analysés	43

Les vidéos du sous-corpus gaming	43
Critères de sélection des extraits.....	44
Tableau récapitulatif des extraits	45
Lecture du tableau.....	45
8.2 Analyse des extraits	46
Fiches d'analyse.....	46
8.3 Synthèse du doublage automatique face aux vidéos gaming.....	71
9. Analyse du sous-corpus de vulgarisation scientifique	75
9.1 Présentation des extraits analysés	75
Les vidéos du sous-corpus de vulgarisation scientifique.....	75
Critères de sélection des extraits.....	76
Tableau récapitulatif des extraits	76
Lecture du tableau.....	77
9.2 Analyse des extraits	78
9.3 Synthèse du doublage automatique face aux vidéos de vulgarisation scientifique	89
PARTIE IV – RÉSULTATS ET DISCUSSION	93
10. Comparaison intergenres	94
10.1 Points de convergence entre gaming et vulgarisation scientifique	94
10.2 Points de divergence	95
10.3 Hiérarchisation des critères selon les genres	97
11. Conclusion générale.....	99
11.1 Synthèse des résultats	99
11.1.1 Par critère	99
11.1.2 Par genre	99
11.2 Retour sur les hypothèses.....	100
11.3 Le voice-over algorithmique.....	102
11.4 Limites de l'analyse	105

11.5 Perspectives.....	106
Bibliographie	108
Filmographie.....	112
Annexes	113

Liste des abréviations et notations

ALST — sigle du projet de recherche sur l’automatisation du voice-over dirigé par Anna Matamala (Universitat Autònoma de Barcelona)

ASR — Automatic Speech Recognition (reconnaissance automatique de la parole)

TA — Traduction automatique

TAN — Traduction automatique neuronale

TER — Translation Edit Rate

TTS — Text-to-Speech (synthèse vocale)

VO — Version originale

WER — Word Error Rate

WSF — World Science Festival

INTRODUCTION GÉNÉRALE

Contexte

Depuis septembre 2025, YouTube propose à ses utilisateurs un système de doublage automatique permettant de visionner dans leur langue les vidéos de créateurs de contenu s'exprimant dans une autre. Le dispositif repose sur une chaîne de traitement entièrement automatisée, alimentée par le modèle Gemini de Google, qui articule reconnaissance automatique de la parole, traduction automatique neuronale et synthèse vocale. Il s'applique rétroactivement aux vidéos existantes éligibles et s'étend à mesure que de nouvelles paires de langues sont prises en charge.

L'ampleur du phénomène le distingue de tout ce que la traduction audiovisuelle a connu jusqu'ici. Là où le doublage professionnel traite un nombre limité de productions sélectionnées, arbitrées par des studios et soumises à des contrôles de qualité, le doublage automatique s'applique indistinctement à un volume de contenus que l'industrie n'aurait jamais pu absorber. Cette généralisation s'accompagne d'une redéfinition implicite de son objet : les contenus concernés ne sont pas des films ni des séries, mais des vidéos de créateurs, dont les conventions communicationnelles diffèrent sensiblement de celles de la fiction audiovisuelle sur laquelle s'est construite la réflexion sur le doublage.

Le choix terminologique opéré par la plateforme mérite en lui-même d'être interrogé. YouTube parle de doublage. Or ce terme désigne, dans les études en traduction audiovisuelle, une pratique définie par un ensemble de standards de qualité précis, portant sur la synchronisation, la crédibilité des dialogues, la cohérence entre image et paroles, la fidélité, la qualité sonore et le jeu d'acteur (Chaume 2012, pp. 15-19). Rien ne garantit *a priori* qu'un dispositif automatisé satisfasse à ces exigences, ni même qu'il vise à y satisfaire. Il est significatif à cet égard que la feuille de route européenne pour les technologies langagières, qui annonçait dès 2012 l'apparition prochaine de systèmes de revoicing automatisés, employait le terme de voice-over et non celui de doublage (Rehm et Uszkoreit 2012, p. 38, cité dans Matamala 2018, p. 75). Le glissement lexical qui s'est opéré depuis n'est pas anodin, puisqu'il conduit à évaluer le dispositif à l'aune d'un modèle dont il ne relève peut-être pas.

État de la recherche

L'automatisation partielle des modes de revoicing n'est pas une préoccupation nouvelle dans la discipline. Chaume signalait déjà, dans son manuel de référence, l'existence de projets visant à traduire automatiquement des dialogues de film à partir d'une transcription obtenue par reconnaissance vocale, et leur reprochait de traiter le texte sans considération pour l'image, au mépris des relations de cohérence et de cohésion entre dialogue et visuel (2012, p. 16). Une décennie plus tard, le projet ALST a étudié plus systématiquement l'implémentation conjointe de ces trois opérations dans le voice-over, en concevant leur automatisation comme un flux de travail dont chaque étape devait rester soumise à une révision humaine (Matamala 2015, cité dans Matamala 2018, p. 75). Le chapitre 3 en détaille le dispositif et les implications.

C'est sur ce point que le dispositif de YouTube marque une rupture, dont le chapitre 3 détaillera les implications : il reprend l'architecture technique envisagée par la recherche en supprimant la boucle de contrôle humain, alors que les travaux d'Ortiz-Boix et Matamala (2016) ont montré que c'est précisément à la mise en voix, et non dans le transfert linguistique, que se creuse l'écart entre traduction humaine et sortie automatique.

Les études en traduction audiovisuelle ont par ailleurs établi que la réussite d'un doublage ne se joue pas au seul niveau verbal. La voix doublée participe à la construction du sens, en véhiculant des informations relatives à l'attitude du locuteur, à ses intentions communicatives et à sa relation avec son public (Bosseaux 2018, pp. 52-54). Elle engage en outre une dimension corporelle : le grain d'une voix porte la trace audible du corps qui la produit (Barthes 1977, p. 179, cité dans Bosseaux 2015, p. 51), et remplacer une voix par une autre revient à déconstruire le lien entre le corps et la voix (Bosseaux 2015, p. 110). Ces acquis n'ont pas encore été confrontés à un dispositif dans lequel la voix cible n'est produite par aucun corps.

Un dernier constat justifie ce travail. La réflexion sur l'oralité préfabriquée s'est construite à partir de corpus de fiction — séries, sitcoms, dessins animés (Baños-Piñero et Chaume 2009 ; Chaume 2012). Les contenus de créateurs relèvent d'un régime discursif différent, où la relation entre l'énonciateur et son audience constitue une part substantielle de ce qui est communiqué, et où les conventions varient fortement d'un genre à l'autre. Ce déplacement d'objet appelle une vérification empirique.

Problématique

La question qui organise ce travail peut dès lors être formulée ainsi : le doublage automatique de YouTube parvient-il à restituer l'acte communicationnel propre aux contenus de créateurs — en termes de synchronisation, d'incarnation vocale et de crédibilité discursive — dans les genres du gaming et de la vulgarisation scientifique ?

Par acte communicationnel, on entendra ici l'ensemble de ce qu'une prise de parole accomplit auprès de son destinataire : établir une relation, manifester une attitude, orienter une interprétation, engager une présence. Cette notion sera précisée au chapitre 2, où l'on montrera qu'elle suppose l'engagement d'un corps dans la voix.

Trois sous-questions en découlent, qui correspondent aux trois axes de la question principale. La synchronisation interroge la correspondance entre la voix produite et ce que montre l'image. L'incarnation vocale interroge la capacité de la voix cible à se donner comme émanant du corps visible à l'écran. La crédibilité discursive interroge la mesure dans laquelle le discours doublé demeure recevable dans la langue cible, au regard des conventions de son genre.

Hypothèses

Trois hypothèses sont soumises à l'épreuve du corpus.

La première pose que le doublage automatique de YouTube ne satisfait pas aux critères qui définissent le doublage dans les études en traduction audiovisuelle, et que les défaillances observées ne relèvent pas d'imperfections ponctuelles susceptibles d'être corrigées par un perfectionnement des modèles, mais de propriétés architecturales du dispositif.

La deuxième pose que ces défaillances s'aggravent à mesure que croissent la densité de l'oralité préfabriquée du discours source et la complexité de sa configuration sonore. Le contraste entre les deux genres retenus, et à l'intérieur de chacun entre des formats plus ou moins planifiés, doit permettre d'en éprouver la validité.

La troisième pose que le dispositif ne relève pas du doublage au sens où l'entendent Chaume, Bosseaux ou Marzà et Chaume. Il produit un substitut vocal standardisé, orienté vers le transfert informatif, dépourvu d'incarnation et laissant subsister une trace audible de la voix

originale. Ces traits l'apparentent au voice-over. On soutiendra toutefois qu'il n'en relève pas davantage, le voice-over étant lui-même régi par des synchronies et des conventions d'authentification que le dispositif ne respecte pas. C'est cette position singulière que l'on proposera de nommer *voice-over algorithmique*, notion élaborée au terme de l'analyse.

Corpus

Le corpus réunit cinq vidéos en anglais, dont le doublage automatique en français a été généré par YouTube. Il se répartit en deux sous-ensembles, retenus pour la distance qui sépare leurs conventions communicationnelles.

Le sous-corpus consacré au gaming comprend trois vidéos. Un guide de démarrage du jeu *Outer Wilds* publié par la chaîne Captain Collins constitue la vidéo principale : un locuteur unique s'adresse à la caméra dans un format tutoriel au débit maîtrisé, configuration qui place le dispositif dans les conditions les plus favorables du sous-corpus. Un extrait de stream de la chaîne Maya, consacré à *Elden Ring*, introduit la pluralité de locuteurs, le jargon spécialisé et la réaction en temps réel. Une séquence de la chaîne Doael, chaîne secondaire du streamer Dantes, consacrée à *League of Legends*, constitue un corpus complémentaire ciblé dont trois extraits seulement sont retenus : cinq locuteurs y interviennent simultanément, dans un contexte de chevauchements constants et de terminologie hyperspécialisée. Quatorze extraits en sont issus.

Le sous-corpus consacré à la vulgarisation scientifique comprend deux vidéos. *Straight Talk on Cosmic Origins*, diffusée par la chaîne World Science Festival, est une conversation en visioconférence entre le physicien Brian Greene et le cosmologiste Andreas Albrecht, retransmise sans montage sur un écran scindé qui maintient les deux intervenants visibles ; le discours y est semi-préparé et dense en marqueurs para-verbaux. Un monologue face caméra de Brian Cox, produit par la chaîne Big Think et consacré aux trous noirs, offre le pendant du tutoriel de Captain Collins : locuteur unique, discours planifié, registre oscillant entre le soutenu et le pédagogique. Sept extraits en sont issus.

L'analyse porte ainsi sur vingt et un extraits, sélectionnés selon une procédure de repérage explicitée au chapitre 6.

Démarche

L'approche adoptée est descriptive et qualitative. Chaque extrait est examiné au moyen d'une grille organisée en cinq catégories de critères — synchronisation, prosodie et expressivité, crédibilité discursive, traduction automatique, relation locuteur-spectateur — selon un protocole en quatre étapes conçu pour limiter les biais liés à une exposition simultanée aux deux versions. Pour chaque critère, la manière dont il se réalise dans la version originale puis dans le doublage est décrite, et l'écart entre les deux est qualifié.

Deux limites doivent être signalées d'emblée. Le doublage automatique évolue au fil des mises à jour des modèles qui l'alimentent, de sorte que les observations rapportées ici correspondent à l'état du dispositif au moment du visionnage. Par ailleurs, l'analyse ne comporte pas d'évaluation par des spectateurs, ce qui la prive d'un accès direct à la réception — une lacune qu'elle partage avec l'essentiel de la recherche sur le voice-over, et sur laquelle on reviendra en conclusion.

Structure du travail

La première partie établit le cadre théorique. Le chapitre 1 définit le doublage comme entreprise d'illusion d'authenticité et le distingue du voice-over. Le chapitre 2 examine la voix comme condition de réussite du doublage et développe la question de sa corporalité. Le chapitre 3 présente le dispositif de YouTube et formule l'hypothèse de son rapprochement avec le voice-over. Le chapitre 4 expose le fonctionnement de la traduction automatique neuronale et ses limites face au discours oral. Le chapitre 5 précise la notion d'oralité préfabriquée et caractérise les deux genres retenus.

La deuxième partie expose la méthodologie : constitution du corpus et sélection des extraits (chapitre 6), grille analytique et protocole (chapitre 7).

La troisième partie présente l'analyse du corpus, consacrée au gaming (chapitre 8) puis à la vulgarisation scientifique (chapitre 9), chaque chapitre se concluant par une synthèse.

La quatrième partie confronte les résultats. Le chapitre 10 compare les deux genres et hiérarchise les critères selon leurs conventions respectives. Le chapitre 11 revient sur les trois hypothèses, déploie la notion de *voice-over algorithmique*, expose les limites de l'analyse et ouvre sur les perspectives qu'elle dégage.

PARTIE I – CADRE THÉORIQUE

1. Le doublage comme illusion d'authenticité

Le doublage est une forme de traduction audiovisuelle qui consiste à remplacer la bande-son originale d'un produit audiovisuel par une autre bande-son où les dialogues ont été enregistrés dans la langue cible. Il ne saurait être réduit à une simple transposition linguistique. Comme le montre Frederic Chaume dans *Audiovisual Translation : Dubbing*, le doublage repose sur une interaction complexe entre langage verbal, image et voix. Chaque genre audiovisuel obéit par ailleurs à des conventions que le public a intégrées, au point que l'absence d'un élément attendu se remarque davantage que sa présence. Chaume emprunte à Lotman la notion de « mécanisme négatif » pour décrire ce phénomène : de même que l'absence de rime dans un genre poétique qui en comporte habituellement produit un effet en soi, l'absence de synchronisation dans un film doublé — là où le public s'attend à la trouver — est perçue comme un défaut et peut compromettre la réception de la production entière (Lotman 1982, p. 125, cité dans Chaume 2012, p. 14). La voix doublée ne se contente donc pas de véhiculer un contenu sémantique équivalent : elle participe pleinement à la construction du sens, incarnant un personnage ou un locuteur dans une situation de communication donnée. La pratique du doublage est à ce titre encadrée par un ensemble de standards de qualité. Chaume en distingue six : la synchronisation acceptable, qui recouvre la synchronisation labiale, la synchronisation cinésique et l'isochronie ; la crédibilité et le réalisme des dialogues ; la cohérence entre image et paroles ; la fidélité de la traduction ; la qualité sonore ; et le jeu d'acteur (2012, pp. 15-19). Ces critères ne relèvent pas uniquement de contraintes techniques, mais conditionnent la capacité de la voix doublée à s'intégrer de manière crédible à l'image et à accomplir sa fonction d'illusion de l'authenticité.

À côté du doublage, la traduction audiovisuelle orale recourt à d'autres modes de transfert, parmi lesquels le voice-over occupe une place spécifique. Díaz Cintas et Orero le définissent comme une technique par laquelle une voix proposant une traduction en langue cible se fait entendre simultanément par-dessus la voix en langue source, dont le volume « est réduit à un niveau bas qui reste audible en arrière-plan pendant que la traduction est lue » (2010, p. 441). Cette superposition obéit à une organisation réglée : l'usage veut que le spectateur entende d'abord la parole originale en clair à l'attaque, que le volume de l'original soit ensuite abaissé pour laisser place à la traduction, et que celle-ci s'achève plusieurs secondes avant la fin de l'énoncé source, le son de l'original étant alors ramené à son niveau normal (ibid., cité dans Matamala 2018, p. 64). Contrairement au doublage, le voice-over ne vise pas à créer une illusion d'authenticité par effacement de la source, mais à la produire par sa conservation partielle.

Franco, Matamala et Orero observent à cet égard que les traits spontanés — hésitations, répétitions, faux départs, anomalies syntaxiques — disparaissent de la version voiced-over pour deux raisons : atteindre l'isochronie propre à ce mode et privilégier la compréhension. Là où le doublage s'efforce de recréer des dialogues crédibles, le voice-over privilégie la transmission intelligible d'un contenu, ce qui suppose d'effacer les traits caractéristiques de la langue parlée spontanée (2010, p. 74, cité dans Matamala 2018, p. 72). Ce mode est de ce fait répandu dans les genres factuels et non fictionnels : documentaires, entretiens, programmes d'actualité.

La distinction entre ces deux modes ne se réduit toutefois pas à leurs procédés techniques. Elle engage ce que chacun décide de faire de la présence de l'étranger dans le produit traduit. Le doublage suppose que le spectateur puisse oublier qu'il regarde une œuvre venue d'ailleurs ; le voice-over lui rappelle constamment le contraire. Díaz Cintas et Orero le formulent en ces termes : « *Whereas dubbing creates the illusion that people on screen speak the same language as the viewer's by hiding the translation act, voiceover is an inescapable reminder of the presence of the Other and the need for translation* » (2010, p. 442). Ce que le doublage s'emploie à masquer, le voice-over l'assume : l'un fait disparaître la traduction derrière la fiction d'une parole immédiatement accessible, l'autre la donne à entendre comme telle, avec la voix source qui continue de courir sous la voix traduite. Les deux modes ne se distinguent donc pas seulement par leur degré d'exigence technique, mais par le régime de visibilité qu'ils imposent à l'opération de traduction elle-même.

Cette différence de finalité ne signifie pas que le voice-over serait dépourvu de contraintes. Matamala en recense quatre, établies d'après Franco, Matamala et Orero ainsi qu'Orero : l'isochronie de voice-over, par laquelle la traduction commence quelques mots après l'original et s'achève avant lui ; la synchronie littérale, qui veut que les mots audibles en début et en fin d'unité, non recouverts par la traduction, soient rendus littéralement, afin que les spectateurs comprenant la langue source puissent rapporter la traduction à l'original ; la synchronisation cinésique, qui accorde la traduction aux mouvements corporels des locuteurs ; et la synchronie d'action, qui l'accorde aux images (2020, pp. 134-135). Le voice-over constitue donc un mode normé, et non un simple relâchement des exigences du doublage. Cette précision sera déterminante pour l'évaluation du dispositif étudié ici.

Dans *The Language of Dubbing : Present Facts and Future Perspectives*, Marzà et Chaume rappellent que la plupart des travaux consacrés au doublage accordent davantage d'importance au réalisme des dialogues qu'à la qualité de la synchronisation labiale : si les gestes, l'intonation

et les dialogues sont crédibles et naturels, le public se montre tolérant envers les défauts de synchronisation qui pourraient apparaître (2009, p. 35). Ce qui importe, résume Whitman-Linsen, c'est « l'impression, la crédibilité de la parole saisie comme un tout » (1992, pp. 54-55, cité dans Marzà et Chaume 2009, p. 35). Caillé formulait déjà cette priorité en 1960: la synchronisation labiale n'est décisive que dans les gros plans, l'essentiel étant de restituer le rythme, la sensibilité, la colère ou la tendresse portés par la voix originale, et « si les voix des acteurs sont judicieusement choisies, si le texte doublé est judicieusement traduit, s'il émeut ou s'il divertit, nous avons réussi » (1960, p. 107, cité dans Bosseaux 2018, p. 54). C'est à lui que l'on doit la formule qui condense cette exigence : le cinéma étant une fabrique d'illusions, le doublage vise « l'illusion d'une illusion » (1960, p. 108), mot que Whitman-Linsen puis Bosseaux (2018, p. 55) ont l'un et l'autre repris à leur compte. Dès lors, les standards du doublage permettent à la voix cible de produire un effet pragmatique et expressif comparable à celui de la version originale. Le doublage n'est pleinement réussi que lorsque la voix cible accomplit un acte communicationnel équivalent à celui de la voix source. Cette conception constitue un cadre de référence indispensable pour interroger les formes émergentes de doublage automatisé.

2. La voix comme vecteur de réussite du doublage

Dans un contexte audiovisuel, le langage ne saurait être appréhendé indépendamment de sa réalisation orale. Contrairement à la traduction écrite, le doublage implique une matérialisation vocale du texte traduit, qui s'inscrit dans une situation de communication où l'image, le corps et la voix interagissent de manière constante. Plusieurs travaux soulignent qu'il est essentiel de préserver les qualités des voix des acteurs originaux (Mingant 2010 ; Bosseaux 2015 et 2018 ; Sanz Ortega 2015, recensés dans Bosseaux 2018, p. 54). La voix originale de l'acteur étant absente de la version doublée, le doubleur doit recréer l'illusion que le corps visible à l'écran et la voix entendue émanent d'une même présence : « dans le doublage, le corps original est séparé de sa voix originale, même si les films doublés donnent l'illusion que la voix et le corps travaillent de concert » (Bosseaux 2018, p. 54 ; formulation reprise de Bosseaux 2015, p. 66). La voix constitue ainsi le pilier du doublage. Comme le rappellent les contributions réunies dans *The Routledge Handbook of Audiovisual Translation* (Pérez-González 2018), le doublage ne résulte pas uniquement du verbal, mais de la combinaison de paramètres linguistiques, prosodiques et visuels. Dans son article *Investigating Dubbing*, Bosseaux souligne que le doublage ne peut être évalué à partir des seuls critères linguistiques, car il dépasse la simple correspondance textuelle (2018, pp. 52-54). Selon elle, la voix doublée participe à la construction du sens au même titre que le texte traduit, dans la mesure où elle véhicule des informations relatives à l'attitude du locuteur, à ses intentions communicatives et à sa relation avec l'audience.

C'est cette dernière propriété qui permet de préciser la notion d'acte communicationnel autour de laquelle s'organise ce travail. On entendra par là non le seul contenu propositionnel d'un énoncé, mais l'ensemble de ce qu'une prise de parole accomplit auprès de son destinataire : établir une relation, manifester une attitude, orienter une interprétation, engager une présence. Un même énoncé peut ainsi accomplir des actes très différents selon la façon dont il est prononcé : « c'est génial » dit sur un ton enthousiaste félicite, dit sur un ton plat constate, dit sur un ton traînant ironise. Le contenu propositionnel est identique dans les trois cas ; l'acte accompli ne l'est pas. La réussite du doublage repose dès lors sur la capacité de la voix cible à accomplir dans la langue d'arrivée un acte équivalent, et non seulement à en restituer le contenu. Cette exigence, on va le voir, engage le corps.

2.1 La corporalité de la voix

Affirmer que la voix constitue le pilier du doublage suppose de préciser ce qui, dans une voix, résiste à la substitution. La réponse tient à ce que l'on peut appeler sa corporalité : le fait qu'une voix ne soit pas un signal détachable de l'organisme qui la produit, mais qu'elle porte les traces audibles de ce corps.

Roland Barthes a proposé pour désigner ce phénomène la notion de grain de la voix. Le grain, écrit-il, est ce qui parvient à l'oreille « d'un même mouvement depuis les profondeurs des cavités, des muscles, des membranes, des cartilages » (Barthes 1977, p. 179, cité dans Bosseaux 2015, p. 51). Élaborée à propos de la voix chantée, la notion distingue les voix qui possèdent un grain de celles qui en sont dépourvues, et permet de rendre compte du plaisir d'écoute que procurent les voix dans lesquelles les traces du corps de l'interprète peuvent être entendues. Le grain désigne ainsi ce que Barthes nomme « l'espace précis de la rencontre entre une langue et une voix » (1977, p. 181). Prolongeant cette analyse, Susan Smith définit la voix comme « le lieu ultime d'intersection et de négociation entre le corps et le langage » (2007, p. 205, cité dans Bosseaux 2015, p. 51).

Cette dimension corporelle explique pourquoi le doublage constitue une opération d'une nature particulière. Comme le formule Bosseaux, « lorsqu'une voix est remplacée par une autre, le lien entre le corps et la voix est déconstruit » (2015, p. 110). Le spectateur d'une version doublée perçoit un corps à l'écran et entend une voix qui n'en provient pas ; l'illusion du doublage consiste à faire tenir ensemble ces deux données contradictoires. Michel Chion, citant Jean Renoir, en tirait déjà une conséquence radicale : « Accepter le doublage, c'est cesser de croire à l'unicité de l'individu » (Chion 1985, p. 74, cité dans Bosseaux 2015, p. 66, et repris dans Bosseaux 2018, p. 54).

La pratique professionnelle du doublage a développé des stratégies pour atténuer cette disjonction. Meyer-Dinkgräfe observe qu'en cas d'incertitude, « la voix de doublage tend à correspondre à l'apparence de l'acteur original davantage qu'à la voix de l'acteur original » (2007, cité dans Bosseaux 2015, p. 71). Il en donne pour exemple la série *Magnum P.I.*, où la voix inhabituellement aiguë de Tom Selleck, peu accordée à son apparence, est remplacée en allemand par la voix nettement plus grave de Norbert Langer. Le choix vocal s'opère donc en fonction du corps visible, et non de la voix source. Cette pratique confirme, par le détour qu'elle

emprunte, que la cohérence entre corps et voix constitue l'enjeu central de l'opération de doublage.

Cette cohérence demeure toutefois instable, y compris dans le doublage professionnel. Bosseaux relève ainsi le cas du comédien français Jean-Pierre Michaël, qui prête sa voix à Brad Pitt, Keanu Reeves, Jude Law, Christian Bale et Ben Affleck : « une voix et un corps uniques dans la version originale donnent lieu à deux, trois ou quatre corps portant tous la même voix dans les différentes versions doublées » (2015, p. 71). Le procédé peut désorienter le spectateur, et il illustre que l'appariement entre un corps et une voix relève d'une convention négociée plutôt que d'une nécessité.

C'est cette précarité que Bosseaux analyse à travers la notion d'inquiétante étrangeté. Le doublage produit selon elle un effet d'inquiétante étrangeté — elle emploie le terme freudien *uncanny* —, « puisque nous entendons une voix qui n'appartient pas au corps de l'acteur que nous voyons jouer » (2015, p. 213). Elle propose de qualifier la voix doublée d'acousmatique, empruntant à Chion (1982) un terme qui désigne à l'origine une voix dont la source n'est pas visible à l'écran. L'application au doublage n'est pas immédiate, puisque le spectateur voit bien un corps émettre la voix qu'il entend ; mais ce corps n'est pas celui dont la voix provient, et le corps d'origine demeure invisible. Bosseaux mobilise à cet égard la description que Mladen Dolar donne de la voix acousmatique : « une voix en quête d'une origine, en quête d'un corps, mais qui, même lorsqu'elle trouve son corps, révèle que cela ne fonctionne pas tout à fait, que la voix n'adhère pas au corps, qu'elle est une excroissance qui ne lui correspond pas » (Dolar 2006, p. 61, cité dans Bosseaux 2015, p. 214). Subsiste ainsi « un écart entre la source de la voix et son résultat auditif, qui ne peut jamais être tout à fait comblé » (Dolar 2006, p. 67). Bosseaux en conclut que la voix doublée peut être décrite comme un fantôme hantant le corps d'un autre acteur, et le doublage lui-même comme une forme de ventriloquie (2015, p. 214).

Ce cadre conceptuel présente un intérêt direct pour l'analyse du doublage automatique, mais il appelle une transposition qui n'est pas immédiate. Chez Bosseaux, la voix doublée est acousmatique parce qu'elle provient d'un corps réel, celui du comédien de doublage, que le spectateur n'aperçoit pas. La disjonction est donc une disjonction entre deux corps. Dans le cas d'une voix de synthèse, aucun corps ne se trouve à la source de l'énonciation. La voix n'est pas en quête d'un corps au sens où l'entend Dolar : elle n'en a jamais eu. Le grain, tel que Barthes le définit, n'y est pas atténué ou transformé, il est structurellement absent, faute de cavités, de

muscles, de membranes et de cartilages dont il puisse porter la trace. La métaphore du fantôme atteint ici sa limite, puisqu'elle suppose une voix ayant quitté un corps.

Cette distinction commande la manière dont les critères d'incarnation vocale seront appliqués au corpus. Il ne s'agira pas d'évaluer si la voix de synthèse parvient à recréer l'illusion d'une présence cohérente entre corps et voix, mais d'observer selon quelles modalités cette illusion se trouve empêchée, et avec quelles conséquences pour la relation entre le créateur de contenu et son audience — une relation qui, dans les genres retenus, repose largement sur la reconnaissance d'une voix comme signature identitaire.

Sur le plan méthodologique, le modèle d'analyse élaboré par Bosseaux fournit par ailleurs un ensemble de paramètres descriptifs directement mobilisables. Reprenant van Leeuwen, elle retient sept critères de qualité vocale, chacun gradué entre deux pôles adjectivaux : la tension, la rugosité, le souffle, l'intensité, le registre de hauteur, le vibrato et la nasalité (2015, p. 103), auxquels s'ajoutent le tempo, le rythme et le volume (2015, pp. 110-111). Le critère du vibrato mérite une mention particulière : van Leeuwen y oppose une voix « unie et sans oscillation — ferme, mais aussi dénuée d'émotion, terne ou plate » à une voix pourvue de grain, c'est-à-dire animée d'une oscillation régulière ou irrégulière (van Leeuwen 1999, p. 134, cité dans Bosseaux 2015, p. 106). Formulée à propos de voix humaines, cette opposition décrit avec une précision inattendue le régime vocal des systèmes de synthèse.

2.2 Prosodie et synchronisation

La prosodie est le premier lieu où se manifeste, à l'oreille du spectateur, cet engagement du corps dans la voix. Les variations d'intonation, d'accentuation, de rythme et de débit permettent de hiérarchiser l'information et de guider l'interprétation du spectateur. Une traduction lexicalement adéquate peut ainsi compromettre le doublage si elle s'accompagne d'une prosodie inadéquate ou artificielle. Bosseaux relève cependant que l'intonation demeure l'un des aspects les moins étudiés du doublage, malgré l'abondance des travaux consacrés à l'oralité préfabriquée, et que Sánchez Mompeán fait à ce jour figure d'exception en ayant abordé la prosodie de façon empirique. Sa méthode consiste à annoter systématiquement, dans la version originale puis dans la version doublée d'un même film, un ensemble de paramètres prosodiques et para-linguistiques — la tonicité, le ton, la tonalité, la direction et l'étendue de la hauteur, l'intensité et le tempo — de manière à mesurer les écarts entre les deux versions plutôt que de les apprécier intuitivement. Ses résultats montrent que les versions doublées sont « souvent

appauvries de la richesse connotative que transmet l'intonation dans la version originale » (Sánchez Mompeán 2016, p. 18, cité dans Bosseaux 2018, p. 55).

Chaume établit pour sa part que la synchronisation constitue l'une des contraintes fondamentales du doublage, et en propose une définition qui en couvre les trois dimensions : « la synchronisation est l'un des traits de la traduction pour le doublage qui consiste à faire correspondre la traduction en langue cible aux mouvements articulatoires et corporels des acteurs et actrices à l'écran, et à garantir que les énoncés et les pauses de la traduction correspondent à ceux du texte source » (2012, p. 68). Trois types en découlent, énoncés à la même page : la synchronie phonétique ou labiale, la synchronisation cinésique et l'isochronie, entendue comme la correspondance entre les énoncés et les pauses. Ces contraintes influencent directement la réalisation vocale du texte traduit et conditionnent la crédibilité de l'illusion audiovisuelle recherchée par le doublage.

3. Le doublage automatique généré par YouTube

Depuis septembre 2025, les utilisateurs de YouTube peuvent regarder des vidéos de créateurs de contenu anglophones dans leur langue grâce à un système de doublage automatique (YouTube Support 2025). Ce dispositif s'inscrit dans une logique d'accessibilité et de diffusion multilingue à grande échelle. Contrairement au doublage professionnel, il repose sur une chaîne de traitement entièrement automatisée alimentée par Gemini, le système d'intelligence artificielle développé par Google, qui combine reconnaissance vocale, traduction automatique neuronale et synthèse vocale. Il vise avant tout à rendre les contenus compréhensibles pour un public élargi, sans intervention humaine dans les phases de traduction, d'adaptation ou d'interprétation vocale. Ce positionnement distingue le doublage automatique de YouTube des pratiques traditionnelles décrites aux chapitres précédents : il ne s'agit pas de produire une illusion audiovisuelle complète, mais de proposer une solution fonctionnelle de transfert linguistique intégrée à l'écosystème de la plateforme.

Cette architecture n'est pas une nouveauté conceptuelle. Une décennie plus tôt, le projet ALST — mené à l'Universitat Autònoma de Barcelona sous la direction d'Anna Matamala — avait étudié l'implémentation conjointe de la reconnaissance vocale, de la traduction automatique et de la synthèse vocale dans le voice-over de documentaires, en concevant ces trois modules comme les composantes d'un flux de travail semi-automatisé dont chacune des trois étapes — génération semi-automatique du script, traduction automatique, restitution vocale de la sortie — devait être « toujours suivie d'une révision ou d'une post-édition humaine » (Matamala 2015, cité dans Matamala 2018, p. 75). Le dispositif de YouTube reprend cette architecture et en supprime la boucle de contrôle. Cette suppression constitue sa différence spécifique avec les systèmes envisagés par la recherche, et son coût avait été mesuré avant même que le dispositif n'existe : l'évaluation à trois niveaux conduite par Ortiz-Boix et Matamala ne relève aucune différence significative entre traduction humaine et post-édition de traduction automatique tant que l'on juge le texte écrit, mais fait apparaître un avantage de la traduction humaine dès lors que ce texte est enregistré, puis soumis aux professionnels du doublage et aux spectateurs (2016, cité dans Matamala 2018, p. 75). L'écart se manifeste au moment de la mise en voix.

L'un des traits saillants du doublage automatique de YouTube réside dans l'utilisation de voix de synthèse, dont les caractéristiques prosodiques et expressives sont largement prédéfinies. Si les avancées en synthèse vocale ont permis d'améliorer la fluidité et

l'intelligibilité du discours, ces voix conservent une expressivité limitée, car le rendu souple et approprié de l'expressivité dans une voix de synthèse demeure « hors d'atteinte », faire sonner une voix joyeuse ou abattue, amicale ou empathique, autoritaire ou hésitante restant « au-delà de ce qu'il est possible de faire aujourd'hui » (Schröder 2009, p. 111). Leur capacité à produire des effets expressifs et interactionnels comparables à ceux d'une voix interprétée s'en trouve réduite.

Il en résulte une forme de neutralisation du discours dans laquelle les émotions, l'emphase et les autres éléments qui confèrent un caractère humain au doublage sont atténués, voire supprimés. Cette neutralisation est particulièrement perceptible dans les formats où le créateur s'adresse directement à la caméra, la relation entre le vidéaste et son spectateur constituant alors un élément central du dispositif communicationnel.

Ces caractéristiques rapprochent le doublage automatique de YouTube non pas du doublage professionnel, mais du voice-over tel que défini par Díaz Cintas et Orero (2010, p. 441) : un transfert linguistique visant avant tout la transmission d'un contenu informatif, sans incarnation vocale ni continuité expressive avec le locuteur à l'écran. La logique fonctionnelle du voice-over — transmission informative plutôt que recreation d'une illusion du réel — présente de fortes similitudes avec le dispositif étudié, et la persistance audible de la voix originale renforce ce rapprochement. On formulera par conséquent l'hypothèse que le doublage automatique de YouTube constitue une forme particulière de revoicing, que ce travail se propose d'appeler voice-over algorithmique et dont la définition ne pourra être établie qu'au terme de l'analyse. Cette hypothèse demande à être éprouvée dans les deux directions : il conviendra de vérifier non seulement que le dispositif ne satisfait pas aux critères du doublage, mais également qu'il ne satisfait pas davantage à ceux du voice-over, qui constituent eux aussi un ensemble de normes contraignantes.

4. Traduction automatique neuronale et oralité préfabriquée

Le dispositif décrit au chapitre précédent repose sur une technologie dont il convient d'exposer le fonctionnement avant d'en évaluer les productions. Ce chapitre retrace l'évolution qui a conduit à l'architecture transformer, décrit la chaîne de traitement mise en œuvre par YouTube, puis examine les limites que rencontre la traduction automatique neuronale face au discours oral et, plus spécifiquement, à l'oralité préfabriquée.

4.1 De la traduction automatique statistique à l'architecture transformer

La traduction automatique a connu, au cours des deux dernières décennies, une transformation paradigmatique profonde. Longtemps dominée par les approches statistiques, fondées sur l'exploitation de corpus parallèles et le calcul de probabilités de correspondance entre séquences de mots, elle a été progressivement remplacée par la traduction automatique neuronale (TAN). Comme l'explique Kenny dans *Machine Translation for Everyone* (2022, p. 40), les systèmes de TAN se distinguent fondamentalement des approches antérieures par leur recours à des réseaux de neurones artificiels constitués de milliers d'unités interconnectées, dont les poids sont ajustés de manière répétitive par exposition à de vastes quantités de données parallèles, de sorte que les prédictions du modèle se rapprochent progressivement d'une réponse jugée correcte. Ces systèmes mobilisent des représentations vectorielles des mots — les *word embeddings* — dont Kenny précise que les mots apparaissant dans des contextes similaires tendent à recevoir des représentations proches lors de l'entraînement (2022, p. 40). Pérez-Ortiz et al. définissent ces *word embeddings* comme des représentations informatisées des mots dotées de propriétés intéressantes, qui, combinées par un mécanisme d'attention, produisent ce que les auteurs nomment des *contextual word embeddings* — des représentations qui encodent non seulement le sens d'un mot, mais aussi son sens en contexte —, constituant un facteur clé dans le fonctionnement des systèmes de TAN (2022, p. 142).

L'avènement de l'architecture transformer, introduite par Vaswani *et al.* dans *Attention Is All You Need* (2017, p. 1), constitue la rupture technique la plus significative de cette évolution. Contrairement aux modèles récurrents antérieurs, qui généraient leurs représentations de manière strictement séquentielle en fonction de l'état précédent à chaque position, le transformer s'affranchit de toute récurrence et repose entièrement sur l'attention pour établir des dépendances globales entre les éléments d'entrée et de sortie (Vaswani *et al.* 2017, p. 2). Cette architecture permet au modèle de pondérer dynamiquement la pertinence de chaque

segment du texte source lors de la génération séquentielle de la traduction : le modèle est dit autorégressif dans la mesure où chaque unité produite est consommée comme entrée supplémentaire pour générer la suivante (Vaswani *et al.* 2017, p. 2). Vaswani *et al.* précisent par ailleurs que le mécanisme d'attention à têtes multiples — *multi-head attention* — permet au modèle de traiter conjointement des informations issues de différents sous-espaces de représentation à différentes positions, plutôt que d'opérer une unique fonction d'attention sur l'ensemble des représentations (2017, pp. 4-5). Cette propriété lui confère une aptitude supérieure à modéliser les dépendances à longue distance au sein d'un énoncé, c'est-à-dire les relations entre éléments éloignés les uns des autres dans la chaîne parlée. On peut en inférer — sans que Vaswani *et al.* traitent eux-mêmes de discours oral — qu'une telle aptitude est utile au traitement de phénomènes fréquents dans la parole spontanée : le réordonnancement syntaxique, lorsque l'ordre des constituants s'écarte de l'ordre canonique ; l'ellipse, lorsqu'un élément nécessaire au sens n'est pas prononcé parce qu'il est récupérable dans le contexte ; et la coréférence, lorsqu'un pronom renvoie à un référent mentionné plusieurs énoncés plus tôt. C'est sur cette architecture que reposent les grands modèles de langage contemporains. Google présente Gemini — le modèle intégré à l'écosystème YouTube pour alimenter son dispositif de doublage automatique — comme reposant sur une architecture de type transformer, sans que le détail de son implémentation soit rendu public.

4.2 Gemini et la chaîne de traitement du doublage automatique YouTube

Le doublage automatique proposé par YouTube depuis septembre 2025 s'appuie sur une chaîne de traitement multimodale articulant trois modules distincts : la reconnaissance automatique de la parole (ASR — Automatic Speech Recognition), la traduction automatique neuronale et la synthèse vocale (Text-to-Speech, TTS). Gemini intervient comme modèle central de cette chaîne en assurant la traduction du contenu transcrit et en participant à l'adaptation du discours cible aux contraintes temporelles nécessaires à sa restitution vocale (YouTube Support, 2025).

Cette architecture multimodale présente des avantages considérables en termes de rapidité de traitement, mais introduit également des points de fragilité à chaque étape de la chaîne. Au niveau de l'ASR en particulier, les erreurs de transcription se répercutent sur les étapes suivantes et peuvent affecter la qualité de la traduction produite. Ruiz et Federico ont démontré, à partir de huit systèmes de reconnaissance vocale évalués sur des conférences TED traduites de l'anglais vers le français dans le cadre de la campagne IWSLT 2013, que le taux d'erreurs de reconnaissance vocale — mesuré par le Word Error Rate (WER), soit la proportion de mots

incorrectement reconnus par rapport à une transcription de référence — prédit de façon statistiquement significative la dégradation de la qualité de traduction, mesurée par le Translation Edit Rate (TER) (2014, p. 267). Il convient de préciser que leur chaîne de traitement reposait sur une traduction automatique statistique et non neuronale, l'étude étant antérieure à la généralisation de cette dernière ; le mécanisme de propagation qu'elle décrit n'en demeure pas moins transposable, puisqu'il tient à l'enchaînement des modules et non à l'architecture du traducteur. Leurs résultats montrent en outre que les différents types d'erreurs de reconnaissance ne contribuent pas de manière équivalente à cette dégradation : les erreurs de substitution portant sur des mots grammaticaux — les mots-outils, par opposition aux mots lexicaux — sont celles dont l'incidence individuelle sur la qualité de la traduction est la plus forte (Ruiz et Federico 2014, p. 271).

Au niveau de la traduction automatique neuronale, une limite plus fondamentale apparaît. Comme le montrent Pérez-Ortiz et al. (2022, p. 146), les systèmes de TAN sont entraînés à partir de paires de textes source et cible produites par des traducteurs humains, et leurs paramètres sont ajustés de manière répétitive afin que leurs prédictions se rapprochent le plus possible de ces exemples. Dans une chaîne de traitement telle que celle de YouTube, la traduction s'effectue à partir d'une transcription du discours oral. Or cette transcription ne conserve qu'imparfaitement certaines dimensions de l'oralité, notamment les variations prosodiques, les hésitations, les autocorrections ou certains marqueurs interactionnels. La traduction produite tend dès lors à traiter le discours comme un texte linguistique autonome plutôt que comme un acte de parole inscrit dans une situation audiovisuelle. Cette caractéristique entre potentiellement en tension avec les exigences du doublage présentées dans les chapitres précédents, qui reposent notamment sur la voix, la prosodie et l'incarnation du discours. Sánchez-Gijón et Kenny soulignent à cet égard que la qualité d'une traduction produite par un système de TAN dépend également de son adéquation à la fonction communicative du texte source, à son registre, à son ton et aux conventions du genre concerné (2022, pp. 88-89).

4.3 Les limites de la TAN face au discours oral et à l'oralité préfabriquée

Le discours oral se distingue du texte écrit par un ensemble de propriétés qui lui confèrent sa texture communicationnelle spécifique : reformulations, faux départs, marqueurs phatiques, adresses directes à l'interlocuteur, variations de registre ou structures syntaxiques incomplètes. Ces caractéristiques, loin de constituer de simples imperfections, participent à la construction

du sens et à l'établissement d'une relation entre locuteur et destinataire. Dans le cadre du doublage audiovisuel, leur traitement adéquat contribue à la crédibilité et à la naturalité du discours cible.

Toutefois, les vidéos analysées dans le travail ne relèvent pas nécessairement d'une oralité spontanée. De nombreux créateurs de contenu mobilisent ce que Chaume (2012, p. 82) désigne sous le terme d'oralité préfabriquée, c'est-à-dire une parole soigneusement élaborée afin de produire un effet de spontanéité. Ce type de discours occupe une position intermédiaire entre l'écrit et l'oral : il est préparé, structuré et parfois scripté, tout en conservant les caractéristiques linguistiques et expressives associées à la communication orale.

Sur YouTube, cette oralité préfabriquée se manifeste sous des formes variées. Les contenus de gaming alternent commentaire en temps réel sur l'action, adresses directes à l'audience et interactions entre joueurs, selon des degrés de planification très variables d'un format à l'autre. Les vidéos de vulgarisation scientifique combinent quant à elles précision terminologique, mise en scène pédagogique et stratégies d'engagement destinées à maintenir l'attention du public. Dans les deux cas, l'impression de spontanéité repose non seulement sur des choix lexicaux et syntaxiques, mais également sur des effets prosodiques, rythmiques et expressifs qui participent à la construction de la relation entre le créateur de contenu et son audience.

Or les systèmes de TAN sont entraînés principalement à partir de textes écrits ou de transcriptions normalisées dans lesquelles une partie des marqueurs de l'oralité a été réduite, corrigée ou standardisée (Pérez-Ortiz *et al.* 2022, p. 146). Ils tendent par conséquent à produire des traductions davantage conformes aux conventions de l'écrit standard qu'aux caractéristiques du discours oral mis en scène. On peut en déduire une série d'effets attendus, que l'analyse du corpus aura précisément pour tâche de vérifier plutôt que de présupposer : un lissage du registre, une réduction des marques d'adresse au public, une homogénéisation stylistique du discours cible. Si ces effets se confirmaient, ils affaibliraient les traits d'oralité qui contribuent à la crédibilité du locuteur et à l'efficacité communicationnelle de la vidéo. Cette conséquence demeure à ce stade une hypothèse dérivée du fonctionnement des systèmes, non un résultat établi.

C'est précisément cette tendance à la neutralisation de l'oralité que l'analyse du corpus cherchera à mettre en évidence. En confrontant les caractéristiques discursives des vidéos sources aux productions du doublage automatique de YouTube, il sera possible d'évaluer dans

quelle mesure le système préserve, transforme ou atténue les marqueurs linguistiques et pragmatiques qui participent à la construction de l'oralité préfabriquée dans différents genres de contenus.

5. Les genres analysés et leurs enjeux pour le doublage automatique

La notion d'oralité préfabriquée, introduite au chapitre précédent, doit à présent être précisée et rendue opératoire. Ce chapitre en expose le cadre d'analyse, caractérise les deux genres retenus au regard de ce cadre, et établit ce que leur comparaison permet de mettre à l'épreuve.

5.1 L'oralité préfabriquée : un cadre d'analyse affiné

Le concept d'oralité préfabriquée, introduit au chapitre 4, peut désormais être précisé. Chaume définit ce mode de discours comme une oralité planifiée, feinte, fabriquée — « *feigned, false, prefabricated* » —, caractéristique de la quasi-totalité des textes audiovisuels de fiction, qu'ils soient originaux ou doublés (2012, p. 82). Le terme même de « préfabriquée » est emprunté au français : c'est Tomaszewicz qui l'introduit dans ce sens (2001, p. 381, cité dans Bosseaux 2018, p. 53). Le dialogue doublé simule en effet la parole spontanée en recourant à des faux départs, des répétitions, des ellipses, des pauses et des interruptions, alors qu'il procède d'un script « écrit pour être dit comme si ce n'était pas écrit » (Gregory et Carroll 1978, cités *ibid.*). Cette oralité ne vise pas à imiter fidèlement la conversation spontanée : l'enjeu, écrivent Baños-Piñero et Chaume, ne réside pas tant dans la tentative d'imiter les conversations spontanées que dans la sélection de traits précis de ce mode de discours qui soient largement admis et reconnus comme tels par le public (2009). Chaume précise de son côté que le doublage procède d'un programme délibéré, et qu'il s'agit moins de reproduire l'ensemble des traits de la langue cible que d'en sélectionner certains (2012, p. 89).

Pour rendre ce concept opérationnel, Chaume mobilise un modèle d'analyse qu'il a lui-même proposé en 2004, testé par Marzà *et al.* sur un corpus de neuf cents minutes de dessins animés doublés puis étendu par Baños-Piñero, et dont la déclinaison pour l'étude du *dubbese* est due à Baños-Piñero et Chaume (2009) ; Chaume en rend compte dans son manuel (2012, p. 87). Ce modèle distingue quatre niveaux linguistiques où se manifeste l'oralité fabriquée : le niveau phonétique et prosodique, le niveau morphologique (la formation des mots et leurs variations grammaticales), le niveau syntaxique (l'organisation des énoncés) et le niveau lexical et sémantique (le choix du vocabulaire) (2012, pp. 88-91). Au niveau prosodique, les traits caractéristiques incluent la clarté d'articulation, l'usage marqué et emphatique de l'intonation, l'allongement de certains sons et l'utilisation de l'intonation comme marqueur de cohésion discursive, c'est-à-dire un moyen de relier les énoncés entre eux par la seule mélodie de la voix (Chaume 2012, p. 89). Au niveau syntaxique, l'oralité fabriquée se manifeste par la

prédominance de phrases courtes et de structures juxtaposées, par des dysfluences — énoncés incomplets, hésitations, reformulations, faux départs, paraphrases expressives —, par l'usage fréquent de marqueurs discursifs (des mots ou expressions qui structurent le discours sans apporter d'information propositionnelle, comme « donc » ou « enfin »), ainsi que par l'ellipse, la deixis et la topicalisation, cette dernière consistant à placer en tête de phrase l'élément dont on va parler afin d'orienter l'attention de l'auditeur (Chaume 2012, pp. 87 et 90 ; Baños-Piñero et Chaume 2009). Les interjections (des mots exclamatifs comme « oh » ou « wow ») et les vocatifs (des formes d'adresse directe à l'interlocuteur, comme « guys ») relèvent quant à eux du niveau lexical et sémantique.

Ce modèle est précieux pour l'analyse du corpus : il permet d'identifier, dans la version originale, les marqueurs précis d'oralité préfabriquée propres à chaque genre, puis d'observer si le doublage automatique de YouTube parvient à les restituer dans ce que Chaume nomme le *dubbese* — la langue propre au doublage, que Pavesi qualifie de « troisième norme », semblable mais non identique au discours oral réel comme aux productions originales de la culture cible (Chaume 2012, p. 87). Un enseignement de ce modèle mérite d'être retenu dès à présent : les porteurs d'oralité du doublage professionnel se concentrent au niveau lexical et sémantique, qui apparaît surchargé par comparaison, tandis qu'ils demeurent limités au niveau phonétique et prosodique et pratiquement inexistant au niveau morphologique (Baños-Piñero et Chaume 2009, section 4 ; l'article a paru dans une revue en ligne dépourvue de pagination). Le doublage humain lui-même n'investit donc que faiblement le niveau prosodique, constat qui devra être gardé à l'esprit au moment d'évaluer les défaillances du dispositif automatique sur ce plan. Les mêmes auteurs relèvent par ailleurs qu'une partie des marqueurs d'oralité est introduite par les acteurs au moment où ils interprètent le script écrit, et non lors de la traduction : l'oralité du doublage se produit donc en partie au stade du jeu.

5.2 Le gaming : une oralité préfabriquée à forte dimension interactionnelle

Les trois vidéos de gaming retenues présentent des configurations communicationnelles distinctes. La première, un guide de démarrage de la chaîne Captain Collins (7 min 10), relève d'un format tutoriel : un locuteur unique s'adresse directement à la caméra dans un cadre maîtrisé, ce qui correspond à une oralité préfabriquée proche du pôle planifié. La deuxième, un extrait de stream de la chaîne Maya (32 min 38), implique plusieurs locuteurs interagissant en temps réel, ce qui relève d'une oralité nettement plus proche du pôle spontané — au sens où Chaume situe le discours de production maison (« *in-house production discourse* », c'est-à-dire

le discours non scénarisé produit directement dans la langue cible, par opposition au discours doublé) plus près du pôle oral que le *dubbese* lui-même (2012, p. 88, Figure 5.1). La troisième, une séquence de la chaîne Doanet, pousse cette configuration à son terme : cinq locuteurs y interviennent simultanément, dans un contexte de chevauchements constants et de terminologie hyperspécialisée.

Dans les trois cas, on attend une densité importante de marqueurs syntaxiques d'oralité fabriquée tels que définis par Chaume : énoncés incomplets, reformulations, répétitions, ainsi qu'un usage fréquent de marqueurs discursifs (Chaume 2012, p. 90), auxquels s'ajoutent au niveau lexical les interjections et les vocatifs caractéristiques du registre — « guys », « oh my god », « let's go » (Chaume 2012, p. 91). Au niveau prosodique, l'intonation emphatique et les variations de débit liées aux réactions en temps réel constituent des traits saillants, particulièrement dans les vidéos de stream.

5.3 La vulgarisation scientifique : entre rigueur terminologique et stratégies d'engagement

Les deux vidéos de vulgarisation scientifique présentent également des configurations contrastées. Le monologue face caméra de Brian Cox, produit par la chaîne Big Think, relève d'un discours entièrement planifié où l'oralité préfabriquée se rapproche du pôle planifié, tout en conservant des marqueurs d'adresse au public destinés à maintenir l'attention. La vidéo du World Science Festival, où deux scientifiques interagissent sur plus d'une heure, permettra d'observer des dynamiques d'interaction comparables à celles du stream de gaming, bien que dans un registre plus formel.

Selon le modèle mobilisé par Chaume, ce type de discours devrait présenter un niveau lexical marqué par une terminologie spécialisée combinée à des stratégies d'expressivité et de créativité lexicale — figures de style, intertextualité, c'est-à-dire références à d'autres textes ou discours — destinées à rendre le contenu accessible (2012, p. 91). Au niveau syntaxique, on peut s'attendre à des structures plus complexes que dans le gaming, mais où la topicalisation et les marqueurs discursifs jouent un rôle similaire dans la gestion du flux informationnel (Chaume 2012, pp. 87 et 90).

5.4 Complémentarité des deux genres pour l'analyse du doublage automatique

La comparaison entre gaming et vulgarisation scientifique permet de tester la capacité du doublage automatique à restituer l'oralité préfabriquée dans deux configurations contrastées. D'une part, le degré de proximité avec le pôle spontané varie significativement : du tutoriel planifié de Captain Collins à l'interaction en temps réel du stream de Maya, en passant par le monologue maîtrisé de Brian Cox jusqu'aux échanges entre les deux scientifiques du World Science Festival. D'autre part, la nature des défis terminologiques diffère fondamentalement : jargon spécifique aux mécaniques de jeu dans un cas, terminologie scientifique dans l'autre — deux domaines où, comme le souligne Chaume, l'analyse doit identifier les traits récurrents propres à chaque registre (2012, p. 91). Une réserve s'impose toutefois sur l'usage du terme de *dubbese* dans la suite de ce travail. Le *dubbese* désigne une norme collective, sédimentée par des décennies de pratique professionnelle et transmise entre traducteurs, adaptateurs et comédiens. Le doublage automatique ne relève d'aucune pratique de ce genre : il n'a ni tradition, ni conventions internes, ni communauté professionnelle pour les faire évoluer. Il ne produit donc pas de *dubbese*, et l'analyse qui suit n'emploiera pas ce terme pour décrire ses productions. Le modèle des quatre niveaux reste en revanche pleinement utile, non pour identifier une norme dans le doublage automatique, mais pour repérer dans la version originale les marqueurs d'oralité dont on observera ensuite le sort.

Cette complémentarité permettra, dans la partie analytique, d'évaluer si les défaillances du doublage automatique relèvent de limites générales du dispositif — affectant uniformément les deux genres — ou de limites spécifiques à certaines configurations communicationnelles, ce qui viendrait nuancer ou renforcer l'hypothèse du *voice-over algorithmique* selon le degré de spontanéité et la densité terminologique du discours source.

PARTIE II – MÉTHODOLOGIE

6. Constitution du corpus

Ce chapitre expose les critères qui ont présidé à la constitution du corpus, présente les cinq vidéos retenues et détaille la procédure de sélection des vingt et un extraits analysés.

6.1 Critères de sélection des vidéos

La constitution du corpus répond à un double impératif. D’une part, les vidéos sélectionnées doivent permettre d’observer les cinq catégories de critères de la grille analytique dans des conditions suffisamment variées pour que les résultats soient représentatifs des enjeux du doublage automatique. D’autre part, elles doivent présenter des configurations communicationnelles contrastées, de façon à ce que la comparaison entre les genres soit analytiquement productive.

Le corpus est structuré en deux ensembles. Le premier, consacré au gaming, comprend trois vidéos : un stream commenté dont six extraits sont retenus, un tutoriel dont cinq extraits sont retenus, et un corpus complémentaire ciblé de trois extraits issus d’une troisième vidéo, choisis pour documenter les conditions les plus défavorables. Le second, consacré à la vulgarisation scientifique, comprend deux vidéos, dont quatre et trois extraits sont respectivement retenus. Aucune vidéo n’est analysée dans son intégralité : le travail porte sur des extraits sélectionnés selon la procédure exposée en 6.3. Le corpus réunit ainsi cinq vidéos et vingt et un extraits. L’ensemble des vidéos sources est en anglais et le doublage automatique analysé est en français.

Il convient de noter que certaines vidéos du corpus ont été publiées avant le lancement officiel du doublage automatique de YouTube en septembre 2025. Le dispositif s’appliquant rétroactivement aux vidéos existantes éligibles, leur inclusion dans le corpus ne pose pas de problème méthodologique, mais implique que le doublage analysé a été généré postérieurement à la publication de la vidéo source.

6.2 Présentation des vidéos

6.2.1 Corpus gaming — Captain Collins (tutoriel)

Chaîne : Captain Collins
Titre : *Outer Wilds Starter Guide* (NO SPOILERS) — How to play this game — Out on Steam

in 2020
 Date de publication : 1er juillet 2020
 Durée : 7 min 10
 URL : <https://www.youtube.com/watch?v=IbtGdABmVpQ>

6.2.2 Corpus gaming — Maya (stream commenté)

Chaîne : Maya
 Titre : Un non-joueur peut-il battre un boss d'*Elden Ring*?
 Date de publication : 14 juillet 2025
 URL : <https://www.youtube.com/watch?v=vG5Zmjkt48U>
 Durée : 32 min 38

6.2.3 Corpus gaming — Doael (corpus complémentaire ciblé)

Chaîne : Doael (chaîne secondaire du streamer Dantes)
 Titre : xQc Lost his MIND after Kadeem said this to him... ft Tyler1 & Tobias Fate
 Date de publication : 8 juillet 2026
 Durée : 1 h 49 min 06
 URL : <https://www.youtube.com/watch?v=wG1kaXTedLI>

6.2.4 Corpus vulgarisation scientifique — Big Think (monologue)

Chaîne : Big Think
 Titre : Brian Cox on how black holes could unlock the mysteries of our universe
 Date de publication : 27 octobre 2023
 Durée : 12 min 51
 URL : <https://www.youtube.com/watch?v=pGsbEd6w7PI>

6.2.5 Corpus vulgarisation scientifique — World Science Festival (entretien)

Chaîne : World Science Festival
 Titre : Straight Talk on Cosmic Origins
 Date de publication : 10 janvier 2025

Durée : 1 h 34 min 54

URL : <https://www.youtube.com/watch?v=Id5iHy59UL8>

La caractérisation communicationnelle de chacune de ces vidéos — format, configuration des locuteurs, degré de planification du discours et enjeux qu’elles posent au doublage automatique — est développée au moment de leur analyse, aux sections 8.1 et 9.1.

6.3 Sélection des extraits

6.3.1 Critères de sélection

Pour chaque vidéo principale, trois à six extraits d’une durée de trente secondes à une minute sont retenus, selon la durée et la densité de la vidéo. Pour les vidéos longues (Maya, World Science Festival), la sélection s’appuie sur un visionnage repérage accéléré. Pour le corpus complémentaire ciblé (Doaenel), trois extraits seulement sont retenus, choisis exclusivement pour leur valeur illustrative sur les critères visés.

La sélection des extraits combine deux logiques. Certains extraits sont ciblés : ils sont choisis parce qu’ils illustrent de façon particulièrement saillante un critère précis de la grille analytique. D’autres sont transversaux : ils permettent d’observer simultanément plusieurs critères en interaction. L’ensemble des extraits retenus pour chaque vidéo doit couvrir, autant que possible, les cinq catégories de la grille.

6.3.2 Procédure de repérage

La sélection s’effectue en quatre étapes. Un premier visionnage de la version originale, sans le doublage, permet de repérer les passages présentant des signaux pertinents pour la grille analytique : forte variation émotionnelle, interpellation directe du spectateur, densité terminologique élevée, débit rapide, geste ou expression faciale marquée, commentaire en temps réel sur l’image, hésitation ou reformulation, référence culturelle ou expression idiomatique, fond sonore important, question rhétorique ou stratégie pédagogique. Une présélection de six à huit passages candidats par vidéo est ensuite constituée à partir des horodatages ainsi notés. Un visionnage du doublage automatique sur chaque passage candidat permet ensuite d’identifier les extraits où des écarts entre version originale et doublage sont observables. La sélection finale retient trois à six extraits par vidéo selon sa durée et sa densité, en veillant à l’équilibre des catégories couvertes.

6.3.3 Limites méthodologiques

Deux limites méthodologiques méritent d'être signalées. D'une part, le doublage automatique de YouTube est susceptible d'évoluer au fil du temps, dans la mesure où les modèles qui l'alimentent sont régulièrement mis à jour par Google. Les analyses présentées dans ce travail correspondent à l'état du doublage au moment du visionnage, et leur reproductibilité exacte ne peut être garantie. D'autre part, la sélection des extraits, bien que guidée par des critères explicites, comporte une part de subjectivité inhérente à tout travail d'analyse qualitative. Cette limite est assumée et constitue une des perspectives d'approfondissement identifiées dans la conclusion.

7. Cadre analytique et protocole d'analyse

L'évaluation du doublage automatique suppose un instrument d'observation explicite. Ce chapitre présente la grille analytique en cinq catégories, précise le niveau auquel s'exerce la description, détaille le protocole appliqué à chaque extrait et expose les conventions de transcription.

7.1 Grille d'analyse : cinq catégories de critères

L'évaluation du doublage automatique de YouTube nécessite un cadre d'analyse structuré, permettant d'observer de manière systématique et reproductible les écarts entre la version originale et la version doublée. La grille analytique adoptée dans ce travail est organisée en cinq catégories de critères, chacune ancrée dans le cadre théorique développé dans la partie précédente. Ces catégories ne sont pas indépendantes les unes des autres — un même phénomène observable peut relever simultanément de plusieurs d'entre elles — mais leur distinction permet de structurer l'analyse et d'identifier avec précision la nature des défaillances ou des réussites observées.

La première catégorie, la synchronisation, regroupe les critères liés à la correspondance entre la voix doublée et l'image. Elle reprend les trois types distingués par Chaume : la synchronisation labiale, définie comme la correspondance entre les mouvements articulatoires visibles à l'écran et les sons produits par la voix doublée ; la synchronisation cinésique, c'est-à-dire l'accord entre la voix doublée et les mouvements corporels, gestes et expressions faciales visibles à l'écran ; et l'isochronie, entendue comme la correspondance entre la durée des énoncés et des pauses de la traduction et celle des énoncés et pauses du texte source (2012, p. 68). À ces trois types s'ajoute la synchronie de personnage, que Chaume définit d'après Whitman-Linsen comme l'adéquation entre la voix entendue et ce que le public attend du locuteur qu'il voit : « en règle générale, un enfant acteur ne peut être doublé par une voix d'homme mûr ; la voix d'une femme doit sonner féminine » (2012, pp. 69-70, cité dans Bosseaux 2018, p. 51). Ce quatrième type n'est pas relevé fiche par fiche, parce qu'il ne se manifeste pas extrait par extrait mais à l'échelle de chaque vidéo : c'est l'attribution des voix par le dispositif, sans considération pour le locuteur visible, qui le met en défaut. Il est donc mobilisé dans les synthèses plutôt que dans les fiches. Les deux synchronies propres au voice-over — la synchronie d'action et la synchronie littérale — ne figurent pas dans la grille, qui est construite d'après les normes du doublage ; elles seront mobilisées au chapitre 11, au moment de confronter le dispositif aux normes de l'autre mode.

La deuxième catégorie, la prosodie et l'expressivité, regroupe les paramètres vocaux qui organisent le discours oral et lui confèrent sa texture communicationnelle. Elle comprend l'intonation, le rythme et le débit, l'emphase et l'accentuation, ainsi que l'expressivité émotionnelle. La description de ces paramètres s'appuie sur les critères de qualité vocale présentés au chapitre 2, que Bosseaux emprunte à van Leeuwen (2015, pp. 103 et 110-111). Comme le souligne Schröder (2009, p. 111), les voix de synthèse modélisent ces paramètres de façon standardisée, ce qui constitue l'une des limites structurelles les plus significatives du doublage automatique.

La troisième catégorie, la crédibilité discursive, évalue dans quelle mesure le discours doublé sonne naturel et crédible dans la langue cible. Elle comprend la naturalité du discours, le registre, la restitution de l'oralité préfabriquée telle que définie par Chaume (2012, p. 82) — marqueurs phatiques, interpellations directes, hésitations construites — et la cohérence audiovisuelle entre contenu verbal et image.

La quatrième catégorie, la traduction automatique, évalue la qualité du transfert linguistique opéré par le système. Elle comprend la fidélité sémantique, la gestion du jargon spécialisé propre à chaque genre, la gestion des culturèmes et des expressions idiomatiques, ainsi que la présence d'erreurs issues de la reconnaissance vocale automatique (ASR), dont Ruiz et Federico ont montré qu'elles constituent un prédicteur statistiquement significatif de la dégradation de la qualité de traduction dans une chaîne automatisée (2014, p. 266).

La cinquième catégorie, la relation locuteur-spectateur, évalue dans quelle mesure le doublage préserve la dimension communicationnelle et relationnelle du discours source. Elle comprend l'incarnation vocale, l'adresse au spectateur et la continuité communicative. Le rattachement de l'incarnation vocale à cette catégorie plutôt qu'à la synchronisation demande à être justifié, car les deux critères concernent le rapport entre la voix et le corps. La synchronisation porte sur une coïncidence mesurable — les lèvres bougent-elles en même temps que les sons sont produits, la durée correspond-elle ? L'incarnation vocale porte sur ce que le spectateur en conclut : reconnaît-il dans cette voix celle de la personne qu'il voit, et par conséquent accepte-t-il de lui parler comme à elle ? C'est un critère relationnel avant d'être technique, et c'est à ce titre qu'il figure ici, au sens développé au chapitre 2 (Bosseaux 2015, pp. 66 et 110). La continuité communicative désigne quant à elle le maintien d'une identité vocale stable et d'un fil énonciatif cohérent d'un tour de parole à l'autre. C'est à ce niveau que se manifeste le plus clairement, ou non, le glissement vers ce que ce travail conceptualise comme un *voice-over algorithmique*, notion élaborée au chapitre 11.

Ces cinq catégories répondent aux trois axes de la question de recherche sans se confondre avec eux. La synchronisation, premier axe, correspond à la première catégorie. L'incarnation vocale, deuxième axe, se distribue entre la deuxième catégorie — puisqu'une voix incarnée se manifeste d'abord par sa prosodie — et la cinquième, qui en évalue l'effet relationnel. La crédibilité discursive, troisième axe, recouvre la troisième catégorie et une part de la quatrième, dans la mesure où une erreur de transfert linguistique compromet la recevabilité du discours autant que sa fidélité. Cette répartition explique que les cinq catégories ne soient pas indépendantes : un même phénomène observable relève fréquemment de plusieurs d'entre elles, et c'est précisément cette interaction que les fiches d'analyse cherchent à documenter.

7.2 Niveau d'analyse et indicateurs observables

La grille analytique adoptée dans ce travail repose sur une logique descriptive plutôt qu'évaluative binaire. Pour chaque critère, il ne s'agit pas de cocher une case indiquant si le doublage est réussi ou non, mais de décrire la valeur observée dans la version originale, puis la valeur observée dans le doublage automatique, et enfin d'estimer l'écart entre les deux.

Cette approche présente plusieurs avantages. Elle permet d'abord de rendre compte de la diversité des configurations observables : un critère peut être parfaitement restitué, partiellement transformé ou entièrement neutralisé, et ces trois situations appellent des formulations différentes qu'une simple case à cocher ne permettrait pas de saisir. Elle permet ensuite de comparer les vidéos entre elles de façon plus fine : deux extraits peuvent présenter un écart de même nature mais d'intensité différente, ce que seule une description qualitative permet de documenter. Elle s'inscrit enfin dans la tradition des études descriptives en traduction audiovisuelle, telle que la mobilisent notamment Baños-Piñero et Chaume (2009), qui privilégie l'identification de tendances et de régularités sur l'émission de jugements normatifs.

Le mot « valeur » demande ici une précision. Il ne s'agit pas d'une mesure chiffrée mais de la caractérisation d'un critère dans un extrait donné : pour l'intonation, par exemple, la valeur relevée dans la version originale pourra être « montée expressive marquée en fin d'énoncé » et celle relevée dans le doublage « contour descendant régulier ». L'écart entre les deux est alors décrit dans les termes de l'observation elle-même, et non codé sur une échelle. Les fiches emploient à cet effet un vocabulaire de gradation constant — écart nul ou absent, partiel, notable, total — dont la fonction est de rendre les extraits comparables entre eux sans prétendre à une quantification que la nature qualitative des observations ne permettrait pas de soutenir.

7.3 Protocole d'analyse

L'analyse de chaque extrait suit un protocole en quatre étapes, conçu pour limiter les biais d'interprétation liés à une exposition simultanée à la version originale et au doublage.

La première étape consiste à visionner l'extrait en version originale, sans le doublage. Ce visionnage a pour objectif de caractériser le discours source selon les critères de la grille : identifier le registre, le débit, les marqueurs d'oralité préfabriquée, les moments d'emphase, les éventuels termes techniques et les émotions perceptibles. Cette étape donne lieu à une transcription partielle des passages les plus pertinents, c'est-à-dire ceux où les phénomènes susceptibles de poser problème au doublage automatique sont les plus saillants.

La deuxième étape consiste à visionner l'extrait en version doublée uniquement, sans le comparer immédiatement à la version originale. Ce visionnage indépendant permet d'évaluer la naturalité du discours doublé en langue cible sans être influencé par la connaissance du texte source, et de noter les premières impressions sur la crédibilité et la fluidité de la voix synthétique.

La troisième étape est la comparaison systématique des deux versions à partir de la grille analytique. Pour chaque critère, la valeur observée dans la version originale et la valeur observée dans le doublage sont décrites, et l'écart est estimé. La transcription partielle réalisée à la première étape est complétée par la transcription des passages correspondants dans le doublage, ce qui permet une confrontation textuelle précise des deux versions sur les passages clés.

La quatrième étape consiste à rédiger le bilan de l'extrait : identification des critères les plus défaillants et des critères partiellement ou entièrement réussis, et formulation des observations les plus révélatrices, en particulier de ce que l'extrait apporte que les autres n'apportent pas. Ce bilan prépare la synthèse par vidéo et la comparaison intergenres développées dans la partie analytique ; l'interprétation conceptuelle des résultats est en revanche réservée au chapitre 11, afin que les fiches restent descriptives.

7.4 Convention de transcription

Les transcriptions partielles adoptent les conventions suivantes :

Les énoncés sont transcrits tels qu'ils sont prononcés, sans correction orthographique ni normalisation syntaxique

Les chevauchements de parole sont signalés par des crochets []

Les pauses brèves sont signalées par (.) et les pauses longues par (..)

Les passages inaudibles ou incompréhensibles sont signalés par (xxx)

Les éléments para-verbaux pertinents (rires, soupirs, changements de débit) sont signalés entre parenthèses (rires), (débit accéléré)

Les phrases ou mots criés sont signalés par des caractères majuscules.

La version originale est transcrite en anglais et la version doublée en français, présentées en regard sur deux colonnes

PARTIE III – ANALYSE DU CORPUS

8. Analyse du sous-corpus gaming

Ce chapitre présente l'analyse des quatorze extraits issus des trois vidéos du sous-corpus gaming. Il expose d'abord les vidéos et les critères de sélection, puis les fiches d'analyse extrait par extrait, avant d'en dégager une synthèse.

8.1 Présentation des extraits analysés

Les vidéos du sous-corpus gaming

Le sous-corpus gaming regroupe trois vidéos représentant des configurations communicationnelles délibérément contrastées, choisies pour tester le doublage automatique dans des conditions différentes.

Maya – Can a non-gamer beat an *Elden Ring* boss ? (YouTube, 14 juillet 2025 ; 32 min 38 s ; 987 021 vues ; 1,25 M abonnés)

Cette vidéo est un extrait de stream Twitch monté par un monteur professionnel (DannyDV) et publié sur YouTube, avec placement de produit. Plusieurs locuteurs y interviennent — Maya, Rosi, Zoil, Sean et Charlie —, mais la parole de Maya domine largement : elle réagit en temps réel à sa partie sur *Elden Ring*, commentant l'action, interpellant ses coéquipiers et s'adressant directement à ses spectateurs. L'oralité préfabriquée y est dense, marquée par de nombreuses interjections, faux départs et marqueurs interactionnels. Le jeu *Elden Ring* fournit par ailleurs un matériau terminologique spécifique — noms de personnages, citations du lore, mécaniques de combat —, dont une partie bénéficie d'une localisation française officielle que le système ignore systématiquement.

Captain Collins – *Outer Wilds Starter Guide* (NO SPOILERS) (YouTube, 1er juillet 2020 ; 7 min 10 s ; 128 005 vues ; 33 700 abonnés)

Cette vidéo a été créée directement pour YouTube sous la forme d'un tutoriel face caméra. Collins s'adresse seul à la caméra dans un registre pédagogique et planifié, illustrant certains passages en diffusant des extraits de ses lives Twitch. Ce format mixte — commentaire explicatif et stream enchâssé — génère une superposition de niveaux énonciatifs que le module ASR ne peut pas distinguer, puisqu'il n'a accès qu'au flux audio. Avec un seul locuteur principal et un débit modéré, la vidéo représente le contexte *a priori* le plus favorable au doublage automatique du sous-corpus gaming.

Doaenel – xQc Lost his MIND after Kadeem said this to him... ft Tyler1 & Tobias Fate (YouTube, 8 juillet 2026 ; 1 h 49 min 06 s ; 119 019 vues ; 709 000 abonnés)

Cette vidéo est un extrait de stream Twitch monté par enigmaxoxoxo et publié sur YouTube. Cinq locuteurs — Dantes (Doaenel), xQc, Tyler1, Tobias Fate et Kadeem — y jouent simultanément à *League of Legends* en vocal. Les chevauchements de parole sont permanents et le jargon du jeu y est particulièrement dense : noms de champions, désignations de mécaniques (CC, knock-up, engage, botlane, etc.), noms d’objets (Black Cleaver, Zhonya’s). Ce format représente les conditions de complexité maximale du sous-corpus et permet de tester les limites architecturales du dispositif face à un contenu pour lequel il n’a manifestement pas été conçu.

Critères de sélection des extraits

Quatorze extraits ont été retenus au total pour le sous-corpus gaming : six pour la vidéo de Maya, cinq pour celle de Captain Collins, et trois pour celle de Doaenel, qui constitue un corpus complémentaire ciblé. La sélection suit deux logiques.

Certains extraits sont ciblés (C dans le tableau) : ils sont retenus parce qu’ils illustrent de façon particulièrement saillante un phénomène précis — une erreur ASR isolée, une substitution phonétique, une défaillance spécifique du module TTS. D’autres sont transversaux (T) : ils permettent d’observer simultanément plusieurs critères de la grille en interaction, ce qui rend compte de la façon dont les défaillances se cumulent plutôt que de se produire isolément. L’ensemble des extraits retenus pour chaque vidéo couvre, autant que possible, les cinq catégories de la grille analytique : synchronisation, prosodie et expressivité, crédibilité discursive, traduction automatique, et relation locuteur-spectateur.

Pour la vidéo de Doaenel, seuls trois extraits ont été retenus, choisis exclusivement pour leur valeur illustrative sur des critères ciblés : l’effondrement total du système face aux chevauchements et au jargon hyperspécialisé, l’incohérence terminologique interne sur un même terme, et la substitution homophonique produisant un énoncé superficiellement cohérent mais sémantiquement faux. Ces trois phénomènes complètent l’analyse sans nécessiter une couverture exhaustive des cinq catégories de la grille.

Tableau récapitulatif des extraits

T = transversal (plusieurs critères en interaction) · C = ciblé (phénomène précis) · Pro = prosodie et expressivité · TA = traduction automatique / ASR · Disc = crédibilité discursive · Rel = relation locuteur-spectateur · Syn = synchronisation

N°	Vidéo	Horodatage	Durée	Type	Phénomènes principaux	Critères
1	Maya	1:00 – 1:15	15 s	T	Neutralisation prosodique de la private joke « blow up » ; voix VO en fond ; voix masculine sur locutrices féminines	Pro · Rel · Disc
2	Maya	4:24 – 4:34	10 s	T	Erreurs ASR sur noms propres du jeu (Dung Eater, Fallen Leaves) ; omission d'une réplique centrale	TA · Disc · Pro
3	Maya	12:55 – 13:01	6 s	T	Suppression du remerciement communautaire (« thank you for the sub ») ; collision VO/doublage sur l'exclamation	Rel · TA · Pro
4	Maya	14:57 – 15:03	6 s	C	Erreur ASR totale : « Don't be dull » transcrit en lettre isolée « D »	TA
5	Maya	15:15 – 15:55	40 s	T	Cinématique Elden Ring : cinq erreurs ASR cumulées dont inversion sémantique (someone → je) et noms propres effacés	TA · Disc · Pro · Rel
6	Maya	18:56 – 19:03	7 s	C	Changement de voix à mi-phrase sur une seule réplique de Maya	Rel · Syn
7	Collins	0:06 – 0:15	9 s	T	Erreur ASR sur le nom de la chaîne (captaincoin) ; pause injustifiée et accélération compensatoire ; Outer Wilds prononcé à la française	TA · Syn · Rel
8	Collins	2:07 – 2:30	23 s	C	Neutralisation totale de l'expressivité spontanée : rires ignorés, interjections plates, cri audible sans équivalent	Pro · Disc
9	Collins	3:01 – 3:21	20 s	T	Discours explicatif planifié : pauses de non-transcription, normalisation de la formule idiomatique, neutralisation du jargon	TA · Disc · Syn
10	Collins	3:49 – 4:12	23 s	T	Fusion de deux niveaux énonciatifs (commentaire + stream enchâssé) ; logs → verrous par substitution phonétique	TA · Disc · Syn
11	Collins	6:42 – 6:51	9 s	C	Rupture prosodique entre le déterminant et le nom (« Un (.) atterrissage »), liaison non réalisée ; fusion énonciative au degré le plus extrême	TA · Disc
12	Doaenel	22:20 – 22:44	24 s	T	Chevauchements constants, jargon LoL (CC, Xayah, Rakan), voix féminine unique pour cinq locuteurs ; cri audible en anglais dans la version doublée	Pro · TA · Rel · Syn · Disc
13	Doaenel	29:45 – 30:05	20 s	C	Incohérence terminologique interne : Zhonya's transcrit Zas puis Zombie dans le même extrait	TA · Disc
14	Doaenel	1:03:02 – 1:03:41	39 s	T	Substitution homophonique (cleaver → clever → je ne suis pas malin) ; inconsistances sur Kassadin et Jax	TA · Disc · Rel

Lecture du tableau

La répartition entre extraits ciblés et transversaux reflète une réalité observée dès les premiers visionnages : les défaillances du doublage automatique se produisent rarement de façon isolée. Sur la majorité des extraits, plusieurs critères de la grille sont simultanément affectés — une erreur ASR sur un nom propre entraîne une perte de crédibilité discursive, qui compromet à son tour la relation locuteur-spectateur. Les extraits transversaux sont donc majoritaires, non par choix arbitraire, mais parce qu'ils reflètent fidèlement le mode de défaillance du système.

Les cinq extraits ciblés (4, 6, 8, 11 et 13) ont été retenus parce qu'ils isolent un phénomène particulièrement net, dont la pureté illustrative justifie une entrée distincte dans le corpus. L'extrait 4 documente l'erreur ASR la plus extrême du sous-corpus gaming — un énoncé entier réduit à une lettre isolée. L'extrait 6 est le seul à documenter un changement de voix à l'intérieur d'une même réplique. L'extrait 8 isole la neutralisation de l'expressivité spontanée dans un contexte sans jargon ni chevauchement. L'extrait 11 documente une rupture prosodique interne au syntagme nominal : la voix de synthèse intercale une pause entre le déterminant et le nom et ne réalise pas la liaison, produisant un « Un (.) atterrissage extrêmement précis » dont la forme écrite est pourtant correcte. L'extrait 13 documente l'incohérence terminologique interne — deux transcriptions différentes pour le même terme dans le même extrait.

La progression entre les trois vidéos n'est pas chronologique mais graduelle en termes de complexité communicationnelle. Collins, avec un locuteur unique et un registre planifié, fournit le niveau de référence le plus favorable. Maya introduit la pluralité de locuteurs, le jargon spécialisé et l'expressivité spontanée. Doael pousse ces trois variables à leur degré maximal simultanément, permettant d'observer l'effondrement du système dans des conditions pour lesquelles son architecture n'a pas été conçue.

8.2 Analyse des extraits

Fiches d'analyse

Vidéo : Maya – *Elden Ring*

M = Maya

Z = Zoil

S = Sean

R = Rosi

N = Narrateur (en jeu)

FICHE D'ANALYSE — EXTRAIT 1	
Vidéo (titre + chaîne)	Un non-joueur peut-il battre un boss d'Elden Ring ? — Maya
Genre	Gaming
Horodatage de l'extrait	1:00 – 1:15
Durée de l'extrait	15 s
Numéro d'extrait	1

Transcription	
VO (anglais)	M-Everybody wants to blow up, Danny can you do that for them, please? Z-Don't let me blow up. M-He's not gonna blow you up. Z-No, please. No, that's so cringe we just left them like that S-Danny, please don't edit anything over him
Doublage automatique (FR)	-Daniel, tout le monde veut exploser. Danny, tu peux faire ça pour eux, s'il te plaît ? -Ne me laisse pas exploser. -Il ne va pas te faire exploser. -Non, s'il te plaît, non c'est tellement gênant, on l'a juste laissé comme ça. -S'il te plaît, Danny, s'il te plaît ne modifie rien par-dessus.
Observations par critère	
Synchronisation	Absence totale de synchronisation labiale : les locuteurs sont visibles à l'écran et leurs mouvements de bouche ne correspondent à aucun moment à la voix doublée. L'isochronie est également problématique — les énoncés doublés semblent lissés, sans respect des pauses, des chevauchements et des reprises qui caractérisent la VO. La synchronisation cinésique est inexistante : les rires, gestes et expressions faciales des intervenants se déroulent à l'écran sans aucun ancrage dans la voix entendue.
Prosodie & expressivité	C'est sans doute la dimension la plus frappante sur cet extrait. La VO est marquée par une forte expressivité collective : rires, tons plaintifs, exclamations, ton complice... qui disparaît entièrement dans le doublage. La voix doublée est monotone, neutre, et produit un discours à débit régulier qui ne reflète ni les montées intonatives de Maya, ni la tonalité suppliante de Rosi. On parle ici d'une neutralisation prosodique totale. Ce que Bosseaux (2018, p. 55) rapporte des versions doublées — appauvries de la richesse connotative que transmet l'intonation dans l'original — s'observe ici à un degré supérieur, puisqu'aucune interprétation vocale n'intervient.
Crédibilité discursive	Trois problèmes s'accumulent ici. D'abord, la naturalité du discours est compromise : les répliques s'enchaînent sans les hésitations, reformulations et chevauchements qui rendent la scène crédible dans la VO. Ensuite, le registre est incohérent — « cringe » devient « gênant », ce qui est fonctionnellement acceptable mais efface la connotation culturelle propre au discours des communautés gaming/YouTube, où « cringe » est un marqueur d'identité. Enfin, l'oralité préfabriquée est partiellement restituée dans le texte (la syntaxe reste informelle), mais complètement écrasée à l'oral par la neutralité de la voix.
Traduction automatique	La fidélité sémantique est globalement maintenue. Le choix de « exploser » pour « blow up » est pertinent dans ce contexte précis, où l'expression renvoie à une blague entre le monteur et Sean.
Relation locuteur–spectateur	L'incarnation vocale est le point le plus saillant : une seule et même voix masculine est utilisée pour l'ensemble des intervenants, y compris pour Maya, locutrice principale de la chaîne, dont la voix est pourtant un élément constitutif de son identité éditoriale. Cette uniformisation vocale rompt brutalement la relation entre le corps visible à l'écran et la voix entendue. Par ailleurs, les voix originales restent audibles en arrière-plan, ce qui crée une dissonance cognitive permanente : le spectateur entend simultanément l'expressivité de la VO et la platitude du doublage, rendant l'illusion de présence impossible à construire.
Bilan de l'extrait	

La séquence transmet son contenu informationnel mais perd sa fonction principale, qui était de créer un moment de complicité humoristique. Aucune erreur de transcription n'est en cause : c'est la seule neutralisation prosodique qui produit la perte.
Réussite : fidélité sémantique, syntaxe informelle préservée à l'écrit. Lacunes : neutralisation prosodique complète, voix masculine sur une locutrice visible, perte du registre sur « cringe », VO audible en fond.

FICHE D'ANALYSE — EXTRAIT 2	
Vidéo (titre + chaîne)	Un non-joueur peut-il battre un boss d'Elden Ring ? — Maya
Genre	Gaming
Horodatage de l'extrait	4:24 – 4:34
Durée de l'extrait	10 s
Numéro d'extrait	2
Transcription	
VO (anglais)	M- The fallen leaves tell a story N- The loathsome Dung Eater. M- What?! R- That's him. M- Did he poop? M- So what- What did- What did, what was all that? R- Yeah.
Doublage automatique (FR)	M- La chute et la fuite racontent une histoire. N- La perte de l'arme. M- Quoi ? R- C'est lui. M- Alors. Qu'est-ce qui a fait quoi. Qu'est-ce qui a fait quoi. Tout ça.
Observations par critère	
Synchronisation	L'isochronie est partiellement maintenue sur les répliques restituées, mais deux omissions massives — « Did he poop? » et « Yeah » — créent des blancs sonores qui ne correspondent à rien de visible à l'écran. Le « What ? ! » original reste audible en fond sonore pendant que la voix doublée prononce « Quoi ? » sans aucune expressivité, ce qui produit une dissonance directement perceptible par le spectateur. La synchronisation cinésique est rompue sur les faux départs de Maya : ses gestes et expressions faciales accompagnent un discours fragmenté et expressif que la voix doublée restitue de façon totalement désarticulée.
Prosodie & expressivité	Les faux départs de Maya (« So what— What did— What did, what was all that? ») sont l'un des passages les plus révélateurs de cet extrait. Dans la VO, ils traduisent une réaction authentique de stupéfaction comique — Maya cherche ses mots en temps réel, et c'est précisément ce désordre expressif qui produit l'effet humoristique. Le doublage restitue la fragmentation syntaxique (« Alors. Qu'est-ce qui a fait quoi. »), mais la voix est plate, le débit régulier, et la ponctuation prosodique absente. La fragmentation devient accidentelle là où elle était construite. L'exclamation « What ? ! » — montée intonative forte, surprise marquée — est aplatie en un simple « Quoi ? » sans relief. La forme est préservée mais le contenu émotionnel disparaît, ce qui illustre l'appauvrissement connotatif que Bosseaux (2018, p. 55) associe à la perte de l'intonation originale.
Crédibilité discursive	Deux problèmes distincts s'accumulent ici. D'abord, l'omission complète de « Did he poop ? » supprime la réplique la plus saillante de la séquence — celle

	qui cristallise la réaction de Maya à la découverte du personnage Dung Eater et qui constitue le pic comique du passage. Un spectateur francophone ne comprend pas pourquoi les autres rient. Ensuite, l'oralité préfabriquée des faux départs n'est pas simplement mal rendue — elle est rendue sans incarnation, ce qui la transforme en dysfonctionnement apparent plutôt qu'en effet rhétorique volontaire.
Traduction automatique	C'est sur ce critère que l'extrait est le plus analytiquement riche. On peut identifier au moins deux erreurs d'ASR majeures. Premièrement, « The loathsome Dung Eater », nom propre d'un personnage dans Elden Ring, prononcé par la narration en jeu est rendu par « La perte de l'arme ». Il s'agit très vraisemblablement d'une erreur de reconnaissance vocale : le système ASR n'a pas identifié le nom propre et a reconstitué phonétiquement une séquence plausible en français à partir de sons anglais mal segmentés. Le résultat est un non sens total qui efface le référent central de la séquence : le spectateur francophone ne sait pas qu'il s'agit d'un nom de personnage, ni pourquoi Maya réagit comme elle le fait. Deuxièmement, « The fallen leaves tell a story » — citation directe du jeu, phrase iconique d'Elden Ring — devient « La chute et la fuite racontent une histoire ». La traduction officielle française du jeu est « Les feuilles mortes racontent une histoire ». Le système a produit une traduction sémantiquement approximative, ignorant l'existence d'une version localisée officielle et effaçant la référence culturelle pour les joueurs francophones. L'omission de « Did he poop? » constitue quant à elle soit une erreur ASR (la phrase n'a pas été reconnue), soit une décision algorithmique de suppression — dans les deux cas, une lacune inacceptable sur le plan de la fidélité sémantique. L'omission de « Yeah » est moins grave mais contribue à la désarticulation du rythme dialogal.
Relation locuteur–spectateur	Contrairement à l'extrait 1, la voix féminine est ici respectée pour Maya, ce qui constitue un point de réussite partiel. Cependant, l'incarnation vocale reste défailante : la voix de Maya dans le doublage ne porte aucune trace de sa personnalité éditoriale — l'enthousiasme naïf et la réactivité spontanée qui définissent son rapport à son audience. La continuité de la relation communicative est rompue par les omissions : le spectateur francophone perd le fil de la réaction de Maya, et donc la dynamique affective qui lie la créatrice à son audience dans ce moment précis.
Bilan de l'extrait	
Ici la compréhension de base est compromise, et non plus seulement la fonction expressive : le spectateur francophone ne peut identifier le Dung Eater ni comprendre l'intensité de la réaction de Maya. C'est le premier extrait où les erreurs ASR sur le jargon rendent la séquence opaque. Réussite : genre vocal féminin respecté, à la différence de l'extrait 1. Lacunes : deux erreurs ASR majeures, omission d'une réplique centrale, perte du référent culturel, collision VO/doublage.	

FICHE D'ANALYSE — EXTRAIT 3	
Vidéo (titre + chaîne)	Un non-joueur peut-il battre un boss d'Elden Ring ? — Maya
Genre	Gaming
Horodatage de l'extrait	12:55 – 13:01
Durée de l'extrait	6 s
Numéro d'extrait	3
Transcription	
VO (anglais)	Z- There you go. M- Hoorray. R- Yay

	M- I did it! Pizza, thank you for the sub. Z- Now, behind you there's a character named Saint Nicholas.
Doublage automatique (FR)	Z- Voilà. M- Hourra. Hey, je l'ai fait. Pizza. Z- Il y a un personnage qui s'appelle Saint Nicholas.
Observations par critère	
Synchronisation	L'absence de synchronisation labiale, déjà documentée sur les extraits précédents, se confirme ici comme une constante du doublage de cette vidéo. Elle n'est donc plus un écart isolé mais une défaillance systémique. L'isochronie est partiellement respectée sur les répliques restituées, mais l'omission de la réplique de Rosi (« Yay ») et de la formule de remerciement de Maya (« thank you for the sub ») introduit des blancs qui ne correspondent à aucune pause visible à l'écran. Plus significativement, la transition entre le « Pizza » de Maya et la réplique de Zoil est inexistante dans le doublage : les deux énoncés s'enchaînent sans respiration, effaçant la structure dialogale de la séquence et produisant un flux continu artificiellement lissé.
Prosodie & expressivité	Le traitement du « Hooray » de Maya constitue le cas le plus marquant de cet extrait sur le plan prosodique. Dans la VO, cette exclamation est portée par une voix féminine reconnaissable, avec une montée intonative forte et une expressivité physique directement liée à un accomplissement dans le jeu. Le système conserve la voix originale de Maya en fond sonore, légèrement atténuée, tout en superposant un « Hourra » robotique et monotone produit par la voix masculine du doublage. Le résultat est une collision sonore entre deux performances simultanées : l'expressivité authentique de la VO et la neutralité du doublage algorithmique se superposent sans jamais fusionner, rendant l'illusion de l'authenticité impossible. Ce phénomène dépasse l'appauvrissement prosodique dont rend compte Bosseaux (2018, p. 55) : il s'agit d'une désynchronisation audible entre deux couches sonores coexistantes, que le spectateur perçoit immédiatement comme artificielle.
Crédibilité discursive	L'omission de « thank you for the sub » supprime une pratique discursive centrale à la culture YouTube : le remerciement en direct des abonnés ou donateurs, adressé nominativement à une personne identifiée (ici « Pizza »). Cette convention, immédiatement reconnaissable pour tout consommateur de contenus gaming, est entièrement effacée dans le doublage. Le spectateur francophone voit Maya prononcer le nom « Pizza » sans contexte ni explication — le doublage restitue le mot isolé sans la réplique qui lui donne son sens. La naturalité du discours est également compromise par l'enchaînement abrupt entre « Pizza » et la réplique de Zoil : sans la formule de remerciement qui l'encadre, le nom propre flotte sans ancrage communicatif, et la transition vers le discours de Zoil perd toute cohérence.
Traduction automatique	Sur le plan de la fidélité sémantique, les répliques restituées sont globalement correctes. « There you go » / « Voilà » et « Now, behind you there's a character named Saint Nicholas » / « Il y a un personnage qui s'appelle Saint Nicholas » ne présentent pas de contresens. On note cependant l'omission du marqueur discursif « Now » en début de réplique chez Zoil, qui dans la VO signale une transition. L'omission de « Yay » (Rosi) est vraisemblablement imputable à une erreur ASR : l'exclamation, courte et phonétiquement simple, n'a pas été transcrite, peut-être masquée par le fond sonore ou par le chevauchement avec la voix de Maya. L'omission de « thank you for the sub » est plus difficile à attribuer à une seule cause : elle peut résulter d'une mauvaise reconnaissance vocale sur

	une réplique prononcée dans l'élan de l'exclamation précédente, ou d'une segmentation algorithmique qui a fusionné les deux énoncés en un seul.
Relation locuteur–spectateur	Cet extrait marque un recul par rapport à l'extrait 2 sur la question de l'incarnation vocale : toutes les voix — masculines et féminines, Zoil, Maya, Rosi — sont rendues par une voix masculine uniforme. La distinction entre locuteurs, partiellement préservée dans l'extrait 2, disparaît ici entièrement. Même en ayant l'image, le spectateur francophone peut difficilement identifier qui parle. Le remerciement nominatif à « Pizza », qui illustre le lien entre Maya et sa communauté, est supprimé sans substitution. Le doublage ne se contente pas d'appauvrir la performance de Maya : il rompt la continuité de la relation communicative entre la créatrice et son audience francophone, reléguant le spectateur à un statut de récepteur passif là où la VO le faisait exister comme membre d'une communauté reconnue.
Bilan de l'extrait	
L'extrait documente la perte de la dimension communautaire : le remerciement adressé à un abonné disparaît entièrement. Il marque en outre une régression vocale par rapport à l'extrait 2, l'uniformisation étant ici totale. Réussite : fidélité sémantique sur les répliques restituées. Lacunes : uniformisation vocale, omission du remerciement communautaire, suppression d'une réplique, collision sonore sur l'exclamation.	

FICHE D'ANALYSE — EXTRAIT 4	
Vidéo (titre + chaîne)	Un non-joueur peut-il battre un boss d'Elden Ring ? — Maya
Genre	Gaming
Horodatage de l'extrait	14:57 – 15:03
Durée de l'extrait	6 s
Numéro d'extrait	4
Transcription	
VO (anglais)	M- Don't be dull. R- Okay, okay. Z- You have to kill him. -Don't be dull. That's awesome
Doublage automatique (FR)	M- D. R- D'accord. Z- Tu dois le tuer M- D. C'est génial.
Observations par critère	
Synchronisation	L'isochronie est approximativement respectée sur les répliques correctement transcrites, mais les deux occurrences de « D » — réduites à une lettre isolée là où Maya prononce une phrase entière — créent des blancs sonores qui ne correspondent à aucune pause observable à l'écran. La synchronisation cinésique est rompue sur ces mêmes occurrences : l'expression de Maya, visible à l'image, accompagne un énoncé complet et expressif que le doublage réduit à une lettre sans signification. La synchronisation labiale, absente de manière constante depuis le premier extrait, n'est pas documentée à nouveau ici comme écart isolé mais confirme le caractère systémique de cette défaillance.

Prosodie & expressivité	La réplique « Don't be dull » est une interpellation directe, vraisemblablement adressée à l'un des participants ou au personnage à l'écran, crie dans la VO. Sa réduction à la lettre « D » dans le doublage supprime tout le sens du contenu. « C'est génial », qui clôt la séquence, est restitué dans une neutralité prosodique totale, sans la montée enthousiaste qui caractérise l'exclamation de Maya dans la VO.
Crédibilité discursive	C'est sur ce critère que l'extrait est le plus analytiquement significatif. La lettre « D » apparaît deux fois dans le doublage, sans contexte ni ancrage communicatif, à des moments où Maya s'exprime de façon parfaitement intelligible dans la VO. Le spectateur francophone se trouve face à un énoncé insignifiant, sans pouvoir en reconstituer le sens. La naturalité du discours est entièrement compromise sur ces occurrences, et la cohérence du dialogue s'effondre : « D. C'est génial. » ne forme pas un enchaînement communicatif recevable. La réplique de Rosi — « Okay, okay » rendu par « D'accord » — est fonctionnellement correcte, bien que la répétition, porteuse d'une nuance d'hésitation dans la VO, soit lissée en un accord neutre dans le doublage.
Traduction automatique	La phrase « Don't be dull » a manifestement été mal segmentée par le module de reconnaissance vocale : seule la première lettre de « Don't » a été reconnue et transcrite. Cette erreur illustre précisément le mécanisme décrit par Ruiz et Federico (2014, p. 266) : une erreur de transcription en amont de la chaîne de traitement se propage sans correction aux étapes suivantes — traduction puis synthèse vocale — jusqu'à produire un résultat incompréhensible dans la version finale. Le système de TAN, recevant « D » comme entrée, n'est pas en mesure de reconstituer l'énoncé source ni d'en inférer le sens à partir du contexte audiovisuel, faute d'accès à l'image ou à la situation de communication. Ruiz et Federico soulignent par ailleurs que les erreurs de substitution portant sur des mots grammaticaux — ici « don't », outil syntaxique fondamental — s'avèrent particulièrement néfastes pour la qualité de la traduction finale (2014, p. 271). Ce cas en constitue une illustration extrême : la perte du mot grammatical entraîne non pas une dégradation, mais une destruction totale de l'énoncé. « You have to kill him » / « Tu dois le tuer » ne présente aucun écart notable.
Relation locuteur–spectateur	L'uniformisation vocale masculine documentée dès l'extrait 3 se confirme ici. L'adresse directe de Maya — « Don't be dull » semble interpeller un interlocuteur précis, participant ou spectateur — est entièrement neutralisée par l'erreur ASR : non seulement la fonction d'adresse disparaît, mais l'existence même de l'énoncé est effacée. Le spectateur francophone ne peut pas reconnaître Maya comme locutrice active sur ces deux occurrences, ce qui rompt la continuité de sa présence communicative dans la séquence.
Bilan de l'extrait	
Extrait ciblé sur l'erreur ASR la plus extrême du sous-corpus : « Don't be dull » réduit à la lettre « D », à deux reprises. Deux énoncés de Maya sont intégralement détruits, sans que le spectateur puisse soupçonner ce qui manque. Réussite : réplique de Zoil correctement traduite. Lacunes : double erreur ASR produisant une lettre isolée, opacité totale sur ces occurrences, neutralisation prosodique de l'exclamation finale.	

FICHE D'ANALYSE — EXTRAIT 5	
Vidéo (titre + chaîne)	Un non-joueur peut-il battre un boss d'Elden Ring ? — Maya
Genre	Gaming
Horodatage de l'extrait	15:15 – 15:55

Durée de l'extrait	40 s
Numéro d'extrait	5
Transcription	
VO (anglais)	M- Cut scene. N- Foul tarnished. In search of the Elden Ring. M- Who's this? R- Gilder, back. N- Emboldened by the flame of ambition. M- No, oh no. I don't- I have to fight this guy? Is this a boss? Why is no one answering? N- Someone must extinguish thy flame. Let it be Margit the Fell. S- You remember the powerpoint? Z- This is our goal.
Doublage automatique (FR)	M- Cinématique. N- Foul terni. À la recherche de l'anneau des anciens. M- Qui est-ce ? R- Gilder, de retour. N- Enhardi par la flamme de l'ambition. M- Non, oh non. Je ne dois pas me battre avec ce type. C'est un boss ? Pourquoi personne ne répond ? N- Je dois éteindre ta flamme. Que ce soit le marché qui tombe S- Oh, toi le powerpoint. Z- C'est notre or.
Observations par critère	
Synchronisation	La longueur de l'extrait et la densité des échanges rendent l'isochronie particulièrement difficile à maintenir pour le système. Les répliques de Maya — notamment la séquence de faux départs exprimant l'inquiétude (« I don't—I have to fight this guy? Is this a boss? Why is no one answering? ») — sont restituées dans un débit régulier et lissé qui ne correspond pas au rythme haché et croissant de l'original. La synchronisation cinésique est rompue de façon particulièrement visible sur les répliques du narrateur, dont la gestualité et la présence à l'image ne trouvent aucun écho dans la voix synthétique du doublage. L'absence de synchronisation labiale, constante depuis le premier extrait, n'est pas redocumentée ici comme écart isolé.
Prosodie & expressivité	C'est sur ce critère que l'extrait est le plus riche analytiquement, et ce à deux niveaux distincts. D'une part, l'intonation du Narrateur dans la cinématique est un marqueur stylistique fort : grave, lente, solennelle, elle construit la menace et l'autorité du personnage de Margit dans un registre archaïque et théâtral, cohérent avec l'esthétique d'Elden Ring. Le doublage restitue cette séquence avec une voix de synthèse neutre, à débit régulier, avec une intonation presque sympathique. La menace est intégralement absorbée par la platitude prosodique. Or les versions doublées sont, selon Sánchez Mompeán, « souvent appauvries de la richesse connotative que transmet l'intonation dans la version originale » (2016, p. 18, cité dans Bosseaux 2018, p. 55) : ici, l'absence de ces choix produit un personnage vidé de toute sa présence visuelle. D'autre part, la réaction de Maya face à la cinématique est construite sur une progression émotionnelle claire dans la VO : surprise (« Who's this ? »), inquiétude croissante (« No, oh no »), puis panique comique (« I have to fight this guy? Is this a boss? Why is no one answering? »). Cette montée en tension, portée prosodiquement par une accélération du débit et une élévation du registre émotionnel, est entièrement neutralisée dans le doublage. La voix cible restitue les mêmes mots à un débit constant et sans inflexion, transformant une réaction authentique en une simple énumération d'interrogations.

Crédibilité discursive	Le registre de Margit constitue le premier point de rupture. L'emploi de la deuxième personne archaïque « thy » dans « extinguish thy flame » signale un registre délibérément élevé et désuet, cohérent avec le ton épique et moyenâgeux d'Elden Ring. Le doublage, s'appuyant sur une traduction automatique incapable de reproduire ce registre, produit un français neutre et moderne qui efface entièrement cette dimension stylistique. La naturalité des faux départs de Maya est également compromise. Comme observé dans l'extrait 2, la fragmentation syntaxique est formellement préservée dans le doublage (« Je ne dois pas me battre avec ce type. C'est un boss ? »), mais sans la texture prosodique qui la rendait authentique dans la VO. Or c'est précisément cette texture qui, selon Chaume (2012, p. 82), distingue l'oralité préfabriquée d'un simple texte mis en voix : la spontanéité simulée repose non sur le texte mais sur sa réalisation vocale.
Traduction automatique	Cet extrait concentre le plus grand nombre d'erreurs traductologiques et ASR documentées dans l'analyse, et mérite un examen détaillé erreur par erreur. « <i>Foul tarnished</i> » → « <i>Foul terni</i> » : le système n'a pas reconnu « tarnished » comme un nom propre — désignation des personnages joueurs dans l'univers d'Elden Ring — et l'a traité comme un adjectif commun, produisant une traduction littérale partiellement incorrecte. La traduction officielle française du jeu est « misérable sans-éclat », formulation qui porte à la fois la connotation méprisante et la logique interne du lore du jeu. Le doublage automatique ignore cette localisation existante et produit un syntagme hybride (« <i>Foul terni</i> ») qui n'appartient ni à la langue du jeu ni à un français courant recevable. « <i>The Elden Ring</i> » → « <i>l'anneau des anciens</i> » : même problème de méconnaissance du référent culturel. Le titre officiel du jeu en français est « Elden Ring » — non traduit — ou, dans certaines occurrences localisées, « l'anneau d'Elden ». « L'anneau des anciens » est une invention du système sans ancrage dans la localisation officielle. « <i>Margit, the Fell</i> » → « <i>le marché qui tombe</i> » : erreur ASR majeure sur un nom propre de personnage. Le système n'a pas identifié « Margit » comme nom propre et a produit une reconnaissance phonétique approximative de l'ensemble du syntagme, générant une séquence sans signification (« le marché qui tombe ») qui efface le nom du boss et détruit la fonction d'annonce dramatique de la réplique. La traduction officielle française, « Margit le Déchu », est ignorée. « <i>You remember the powerpoint?</i> » → « <i>Oh, toi le powerpoint</i> » : erreur ASR sur la structure syntaxique de la question. La phrase interrogative n'a pas été reconnue comme telle, et « you remember » a été réduit à une apostrophe directe (« toi »), produisant un énoncé incompréhensible et incomplet. « <i>This is our goal</i> » → « <i>C'est notre or</i> » : confusion phonétique entre « goal » et « gold » au niveau de l'ASR — deux termes proches acoustiquement, dont le premier, « goal » (objectif), est substitué par « gold » (or). L'erreur est donc ASR et non traductologique : la traduction de « gold » par « or » est juste, mais le mot transcrit est faux. Le résultat est une réplique sémantiquement incohérente dans le contexte de la séquence.
Relation locuteur–spectateur	L'instabilité vocale de Maya, déjà documentée dans les extraits précédents, se confirme ici : une voix masculine lui est par moments attribuée, rompant la continuité de son incarnation vocale. Cette incohérence est particulièrement perceptible dans un extrait où Maya est le locuteur le plus actif et où ses réactions constituent le fil conducteur de la séquence pour le spectateur. L'incarnation vocale du Narrateur est entièrement manquée. Dans la VO, cette voix possède une identité forte — grave, lente, menaçante — qui construit la présence du personnage de Margit avant même que celui-ci apparaisse à l'écran. Dans le doublage, elle est rendue par la même voix synthétique neutre que tous les autres locuteurs, effaçant toute distinction entre personnages et détruisant l'effet de contraste entre la menace du boss et la réaction comique de Maya.
Bilan de l'extrait	

L'extrait le plus dégradé de la vidéo de Maya : cinq erreurs ASR distinctes en quarante secondes, dont une inversion sémantique (« someone » rendu « je ») et deux effacements de référents pourtant localisés officiellement en français. Il documente que le système ne mobilise aucune ressource terminologique propre au jeu. Réussite : répliques courtes et syntaxiquement simples correctement restituées. Lacunes : cinq erreurs ASR dont deux produisant des énoncés dépourvus de sens, neutralisation prosodique du Narrateur, instabilité du genre vocal, perte du registre archaïque de Margit.	
Ancrage théorique	Cet extrait constitue le cas le plus documenté de propagation en cascade d'erreurs ASR dans l'analyse, confirmant le mécanisme décrit par Ruiz et Federico (2014, p. 266) : les erreurs de transcription en amont de la chaîne de traitement se répercutent sans correction sur les étapes suivantes, jusqu'à produire des résultats communicativement nuls. Il illustre en outre une limite fondamentale des systèmes de TAN soulignée par Pérez-Ortiz et al. (2022, p. 146) : entraîné à partir de textes écrits normalisés, le système est incapable de mobiliser la connaissance du contexte audiovisuel, du lore du jeu ou des localisations officielles existantes pour corriger ses propres erreurs. Enfin, la neutralisation prosodique simultanée du Narrateur et de Maya prive la séquence de ce que Marzà et Chaume placent au premier rang des conditions de réussite du doublage : le réalisme et la crédibilité du dialogue, qui priment sur la conformité formelle (2009, p. 35).

FICHE D'ANALYSE — EXTRAIT 6	
Vidéo (titre + chaîne)	Un non-joueur peut-il battre un boss d'Elden Ring ? — Maya
Genre	Gaming
Horodatage de l'extrait	18:56 – 19:03
Durée de l'extrait	7 s
Numéro d'extrait	6
Transcription	
VO (anglais)	R- Okay, Maya, you're doing way better. But, you're so far but close. M- What did you just say?
Doublage automatique (FR)	R- D'accord, Maya, tu t'en sors beaucoup mieux, mais tu es si loin que tu es proche. M- Qu'est-ce que tu viens de dire ?
Observations par critère	
Synchronisation	Aucun écart notable sur ce critère au-delà des défaillances systémiques déjà documentées. L'isochronie est approximativement respectée sur les deux répliques.
Prosodie & expressivité	La réplique de Rosi, dans la VO, est portée par un ton légèrement amusé, conscient du paradoxe de la formule qu'il emploie. Cette nuance disparaît dans le doublage, restitué sur un débit neutre et régulier. La réaction de Maya — « What did you just say? » — exprime dans la VO une stupéfaction comique face à la contradiction apparente de la phrase de Rosi. Sa restitution en français est correcte sur le plan sémantique mais totalement plate prosodiquement, sans la montée intonative qui lui donne sa fonction de réplique humoristique.
Crédibilité discursive	La réplique de Rosi concentre le problème principal de l'extrait. « You're so far but close » est une formule idiomatique exprimant un paradoxe volontaire — tu es encore loin mais tu progresses, tu t'approches. La conjonction « but » marque un contraste délibéré entre deux états opposés, ce qui constitue

	précisément l'effet rhétorique et comique de la phrase. Le doublage produit « tu es si loin que tu es proche », où « que » introduit non un contraste mais une conséquence, transformant un paradoxe assumé en un énoncé logiquement incohérent. La formule perd son humour et sa lisibilité : là où Rosi jouait sur une tension rhétorique, le doublage produit une phrase qui semble mal construite. C'est un cas typique de ce que Pérez-Ortiz et al. (2022, p. 146) identifient comme une limite structurelle des systèmes de TAN : entraînés sur des textes écrits normalisés, ils tendent à traiter les constructions idiomatiques orales comme des séquences littérales, en perdant la fonction pragmatique de l'énoncé.
Traduction automatique	Aucune erreur ASR détectable.
Relation locuteur–spectateur	Cet extrait introduit un phénomène inédit dans l'analyse : un changement de voix à mi-phrase sur une même réplique. La première partie de la réplique de Maya — « Qu'est-ce que » — est rendue par une voix masculine, avant que la voix ne bascule vers une voix féminine pour « tu viens de dire ». Ce dysfonctionnement, distinct des erreurs d'attribution vocale documentées dans les extraits précédents, signale une instabilité du module de synthèse vocale au niveau intra-énonciatif : le système ne maintient pas une identité vocale cohérente sur la durée d'une même réplique. Comme le souligne Schröder (2009, p. 111), les voix de synthèse conservent une expressivité prédéfinie et limitée — mais cet extrait révèle une fragilité supplémentaire : même la cohérence vocale de base, sur laquelle repose toute incarnation du locuteur selon Bosseaux (2015, p. 66), peut être compromise à l'intérieur d'un énoncé unique.
Bilan de l'extrait	
Extrait ciblé sur un phénomène unique dans le corpus : la voix change au milieu d'une même réplique de Maya. L'instabilité vocale ne se situe donc pas seulement entre locuteurs mais à l'intérieur d'un énoncé. Réussite : aucune erreur ASR, isochronie respectée. Lacunes : changement de voix intra-énonciatif, perte de la valeur idiomatique et comique de « so far but close » par traduction littérale, platitude prosodique des deux répliques.	
Ancrage théorique	Le changement de voix à mi-phrase constitue le cas le plus radical d'instabilité d'incarnation vocale documenté dans l'analyse. Là où les extraits précédents montraient une attribution vocale incorrecte mais stable sur la durée d'une réplique, cet extrait révèle que la cohérence vocale peut être rompue à l'intérieur même d'un énoncé. Ce phénomène illustre bien l'écart entre le doublage automatique et professionnel.

Outer Wilds - Captain Collins

FICHE D'ANALYSE — EXTRAIT 7	
Vidéo (titre + chaîne)	Outer Wilds Starter Guide (NO SPOILERS) — Captain Collins
Genre	Gaming
Horodatage de l'extrait	0:06 – 0:15
Durée de l'extrait	9 s
Numéro d'extrait	7
Transcription	
VO (anglais)	Today with a beautiful piece of art, « Outer Wilds ». And I will show you today 10 tips that will give you an anti-frustration guide into the Outer Wilds. So let's go. Before we start, a quick introduction to myself. I'm Captain Collins. I'm a Twitch streamer and YouTube content creator. I stream from Monday to

	Thursday, 7:00 p.m. Central European time on Twitch dot TV slash Captain Collins.
Doublage automatique (FR)	Aujourd'hui, je vous présente une belle œuvre d'art, Outer Wilds (...) Je vais vous montrer dix conseils qui vous aideront à surmonter vos frustrations dans Outer Wilds. Alors, c'est parti (débit accéléré) Avant de commencer, je vais me présenter rapidement. Je suis le Capitaine Collins. Je suis un streamer Twitch et un créateur de contenus YouTube. Je diffuse du lundi et jeudi à dix-neuf heures, heure d'Europe centrale sur Twitch point TV barre oblique captaincoin.
Observations par critère	
Synchronisation	La pause après « Outer Wilds » dans le doublage est le premier écart notable. Captain Collins est visible à l'écran et continue de parler pendant ce blanc sonore — l'isochronie est donc rompue à cet endroit précis, le spectateur voyant des mouvements labiaux sans voix correspondante. Ce n'est pas une pause présente dans la VO : elle résulte vraisemblablement d'une segmentation algorithmique qui a traité « Outer Wilds » comme une fin d'unité prosodique et introduit un blanc de transition. L'accélération soudaine du débit lors de la présentation de Collins illustre un phénomène différent mais lié : le système, contraint de compenser le blanc introduit précédemment, accélère mécaniquement pour rester aligné avec le flux de la VO. Cette correction algorithmique est perceptible précisément parce qu'elle est brutale — le débit passe d'un rythme lent à une accélération sans transition naturelle, ce qui produit un effet de rupture que l'isochronie, censée garantir une continuité fluide entre les énoncés, devrait éviter.
Prosodie & expressivité	C'est le critère le plus révélateur de l'extrait. Dans la VO, Captain Collins adopte un débit délibérément lent et posé sur l'introduction du jeu — une stratégie rhétorique visant à donner du poids à la qualification d'« œuvre d'art » —, avant de s'accélérer naturellement lors de la présentation de ses coordonnées, séquence qu'il récite de mémoire et à un rythme soutenu. Ce contraste de débit est intentionnel et participe à la structuration prosodique du discours. Le doublage restitue la lenteur du premier segment, puis l'accélération du second — mais mécaniquement. Comme le souligne Schröder (2009, p. 111), les voix de synthèse reproduisent des variations de débit de façon prédéfinie, sans la progressivité et la texture corporelle qui rendent l'accélération humaine naturelle. Le résultat est une accélération reconnaissable comme telle mais perçue comme artificielle : le spectateur entend que le système essaie de suivre, pas que le locuteur s'emballe. Par ailleurs, la prononciation « Outer Wilds » — le titre du jeu rendu avec une phonologie française — signale que le module TTS ne dispose d'aucune règle de prononciation spécifique pour les titres anglophones non traduits. C'est une forme d'erreur distincte des erreurs ASR documentées sur la vidéo de Maya : ici, la transcription est correcte, mais la restitution vocale est fautive.
Crédibilité discursive	Le doublage restitue correctement la structure de la séquence de présentation du vidéaste (nom, plateforme, horaires), ce qui constitue une réussite partielle sur ce critère. Cependant, l'accélération robotique décrite plus haut détruit précisément l'effet de familiarité que cette formule est censée produire. L'oralité préfabriquée fonctionne parce qu'elle sonne <i>naturellement</i> répétée, pas mécaniquement récitée.
Traduction automatique	L'erreur la plus significative de l'extrait est la transcription « captaincoin » pour « Captain Collins ». Il s'agit d'une erreur ASR sur le nom propre du créateur — vraisemblablement due à la vitesse du débit à ce moment précis, combinée à la relative similarité phonétique entre « Collins » et « coin » dans un contexte d'énonciation rapide. Conformément au mécanisme décrit par Ruiz et Federico (2014, p. 266), l'erreur de transcription est transmise telle

	quelle au module TTS, qui restitue « captaincoin » comme si c'était un mot français ordinaire. La conséquence est particulièrement grave dans ce contexte précis : le nom de la chaîne Twitch, que Collins cite pour que son audience puisse le suivre, est rendu incorrect. Un spectateur francophone qui ne connaît pas le créateur et chercherait sa chaîne à partir de cette information serait conduit sur une adresse inexistante. Par ailleurs, la traduction « Twitch point TV barre oblique » pour « Twitch dot TV slash » est fonctionnellement correcte — le système traduit les éléments de l'URL en français standard, ce qui est une stratégie cohérente — mais elle sonne étrange à l'oral et renforce l'impression d'un discours artificiel.
Relation locuteur–spectateur	L'erreur sur le nom de la chaîne touche directement la fonction relationnelle de la séquence. Donner son URL Twitch, c'est un acte d'adresse au spectateur — une invitation à rejoindre la communauté du créateur. En déformant « Captain Collins » en « captaincoin », le doublage ne produit pas seulement une erreur informationnelle : il compromet un acte de parole au sens de la théorie des actes de langage (Austin 1962), c'est-à-dire un énoncé dont l'efficacité dépend entièrement de sa précision : nommer correctement le créateur est la condition pour que l'adresse à son audience fonctionne. La relation entre le créateur et son audience francophone est compromise dès les premières secondes de la vidéo.
Bilan de l'extrait	
Premier extrait de Captain Collins, il montre que les défaillances apparaissent sur un contenu pourtant simple — locuteur unique, débit lent, aucun jargon dense. Cela indique que les limites du système tiennent moins à la complexité du contenu qu'à des contraintes d'architecture. Réussite : structure informative correctement traduite. Lacunes : pause injustifiée après « Outer Wilds », accélération robotique, phonologie française sur un titre anglais, erreur ASR sur le nom de la chaîne.	

FICHE D'ANALYSE — EXTRAIT 8	
Vidéo (titre + chaîne)	Outer Wilds Starter Guide (NO SPOILERS) — Captain Collins
Genre	Gaming
Horodatage de l'extrait	2:07 – 2:30
Durée de l'extrait	23 s
Numéro d'extrait	8
Transcription	
VO (anglais)	Wooh. Eh what, on one of these pools? (rires) Wait, what the freak? Uh oh, aaah, aaah. (rires). Can we start that again? Oh yeah, reset.
Doublage automatique (FR)	Oh. Euh. Qu'est-ce qu'il se passe sur l'une de ces piscines. Attendez, c'est quoi ce bordel. Ouh. Ah. Ah. On peut recommencer, oh oui, réinitialiser.
Observations par critère	
Synchronisation	Collins est visible à l'écran pendant toute la séquence, réagissant physiquement à ce qui se passe dans le jeu — rires, posture, expressions faciales. La synchronisation cinésique est entièrement rompue : les éclats de rire visibles ne sont accompagnés d'aucune trace vocale équivalente dans le

	doublage, qui les ignore complètement. Par ailleurs, à un moment où Collins crie de surprise, sa voix originale reste audible dans la bande son sans que le doublage ne la reprenne — créant un blanc entre les deux couches sonores qui expose directement le fonctionnement du système au spectateur.
Prosodie & expressivité	C'est le critère central de cet extrait, et le plus révélateur. La séquence est construite sur une progression émotionnelle dans la VO : surprise (« Wooh »), curiosité (« Eh what, on one of these pools? »), panique (« Wait, what the freak? Uh oh, aaah, aaah »)... Chaque palier est distinct, porté prosodiquement par des montées intonatives, des allongements vocaliques et un débit qui s'emballe avant de se stabiliser. Dans le doublage, cette progression est entièrement absente. Les interjections — « Oh », « Euh », « Ouh », « Ah », « Ah » — sont énoncées sur un débit régulier et sans inflexion, comme une liste lue à voix haute. La panique comique devient un énoncé plat. Les rires, qui dans la VO ponctuent et amplifient chaque pic émotionnel, n'ont aucun équivalent dans le doublage — ils sont simplement ignorés. Ce phénomène illustre précisément ce que Schröder (2009, p. 111) identifie comme limite fondamentale de la synthèse vocale : les variations pragmatiques et émotionnelles de la parole humaine restent impossibles à modéliser de façon fine. Une voix de synthèse peut reproduire une intonation montante ou un débit accéléré de façon isolée, mais elle ne peut pas incarner la texture d'une réaction spontanée.
Crédibilité discursive	Les interjections constituent le nœud du problème sur ce critère. Dans la VO, « Wooh », « Uh oh », « aaah » sont des marqueurs interactionnels qui signalent à l'audience l'état émotionnel du créateur, et qui peuvent construire une relation de complicité entre lui et ses spectateurs. Chaume classe les interjections parmi les marqueurs lexicaux de l'oralité préfabriquée (2012, p. 91) — des éléments qui, bien que formellement simples, jouent un rôle déterminant dans la construction de l'authenticité du discours. Le doublage les restitue graphiquement (« Oh », « Ah », « Ouh »), mais sans la réalisation vocale qui leur donne leur fonction. Le résultat est un discours formellement correct et communicativement vide. La traduction de « reset » par « réinitialiser » mérite également une attention. Dans la VO, Collins utilise un terme du vocabulaire gaming courant — court, direct, immédiatement reconnaissable. « Réinitialiser » est son équivalent dictionnaire exact, mais dans un registre nettement plus formel et plus long, qui contraste avec la décontraction de l'énoncé original. C'est un exemple typique de la tendance des systèmes de TAN à produire des traductions conformes aux conventions de l'écrit standard plutôt qu'au registre oral du discours source (Pérez-Ortiz et al. 2022, p. 146).
Traduction automatique	Sur le plan purement sémantique, la traduction est globalement correcte. « What the freak » est rendu par « c'est quoi ce bordel » — choix fonctionnellement pertinent sur le registre informel, même s'il constitue un glissement vers un niveau de langue légèrement plus vulgaire que l'euphémisme original. La structure des répliques restituées est lisible. Aucune erreur ASR majeure n'est détectable sur cet extrait — les erreurs relèvent ici non pas de la transcription mais de la restitution vocale.
Relation locuteur–spectateur	L'effacement des rires est la défaillance la plus significative sur ce critère. Dans les contenus gaming de type tutoriel, les moments où le créateur rit de ses propres erreurs servent une fonction relationnelle précise : ils humanisent le créateur, montrent qu'il est lui-même soumis aux aléas du jeu, et invitent l'audience à partager ce moment de légèreté. C'est une stratégie d'engagement constitutive du genre, qui relève des vocatifs et marqueurs interactionnels que Chaume range parmi les traits de l'oralité préfabriquée (2012, pp. 90-91), ici transposés aux conventions propres à YouTube. En ignorant les rires, le doublage ne produit pas seulement une omission prosodique : il rompt la dynamique relationnelle de la séquence, transformant un moment de connivence en simple succession d'énoncés.

	Le cri de Collins, audible en fond sonore sans être repris par le doublage, aggrave cet effet : le spectateur entend une expressivité authentique derrière une voix plate, ce qui accentue la distance entre les deux couches sonores plutôt que de la masquer.
Bilan de l'extrait	
Extrait ciblé : les défaillances ne tiennent ici ni à la transcription ni au jargon, mais à la seule incapacité de la voix de synthèse à incarner une réaction spontanée. Le contenu est simple, le débit modéré, et le doublage échoue malgré tout, parce que ce qu'il doit restituer n'est pas de l'information mais de l'expressivité. Réussite : traduction sémantique correcte, registre informel partiellement maintenu. Lacunes : neutralisation totale des interjections et des rires, cri original audible sans équivalent, décalage de registre sur « réinitialiser ».	

FICHE D'ANALYSE — EXTRAIT 9	
Vidéo (titre + chaîne)	Outer Wilds Starter Guide (NO SPOILERS) — Captain Collins
Genre	Gaming
Horodatage de l'extrait	3:01 – 3:21
Durée de l'extrait	20 s
Numéro d'extrait	9
Transcription	
VO (anglais)	As you can see on top it says [Ah men] abort autopilot. So it's the same button to start and to abort the autopilot. We're sticking still kind of to the ship, but this is a more general rule of thumbs there in this game. Be aware of gravity also in late game.
Doublage automatique (FR)	Comme vous pouvez le voir en haut, il est indiqué « abandonnez le pilote automatique ». C'est donc le même bouton pour démarrer et pour abandonner le pilote automatique. (...) Nous restons un peu sur le vaisseau, mais c'est une règle générale dans ce jeu. (...) Soyez conscients de la gravité également en fin de partie.
Observations par critère	
Synchronisation	Les deux pauses (...) signalées pendant lesquelles Collins parle sans que le doublage n'intervienne constituent le principal écart de synchronisation de l'extrait. Le spectateur voit Collins s'exprimer à l'écran sans voix correspondante — l'isochronie est rompue sur ces deux passages, dont la cause est vraisemblablement une non-reconnaissance des segments concernés par le module ASR, peut-être due à un débit plus rapide ou à une articulation moins nette à ces moments précis. Ces lacunes confirment un phénomène déjà documenté sur les extraits précédents : le système ne produit pas de traitement continu du flux audio, mais segmente ce flux en unités qu'il traite de façon discontinue, laissant des blancs là où la transcription échoue. Le traitement du [Ah men] est à cet égard un cas particulier. Cette exclamation, issue d'une couche sonore superposée — un extrait de stream diffusé simultanément dans la vidéo —, est absente du doublage. Deux lectures sont possibles : soit le module ASR a correctement identifié ce son comme un fond

	sonore non prioritaire et l'a ignoré, ce qui constituerait une forme de réussite partielle ; soit il ne l'a simplement pas transcrit faute de le distinguer du discours principal. Dans les deux cas, le résultat pratique est le même — l'absence de traitement — mais les implications sur la capacité du système à gérer des couches audio superposées sont différentes et difficilement tranchables sans accès aux métadonnées de transcription.
Prosodie & expressivité	L'extrait relève d'un registre explicatif et pédagogique — le débit de Collins est posé, l'intonation structurée autour des informations à transmettre. Les variations prosodiques sont moins spectaculaires que dans l'extrait précédent, ce qui limite les défaillances sur ce critère. Le doublage restitue globalement un débit équivalent sur les segments transcrits, sans accélération robotique notable. Ce contexte de discours planifié constitue le terrain le plus favorable au doublage automatique — ce qui rend d'autant plus significatif que des lacunes persistent malgré tout.
Crédibilité discursive	Trois écarts de registre méritent d'être distingués. « Abort autopilot », bref et technique, devient « abandonnez le pilote automatique » : le verbe introduit une nuance d'échec absente de l'original et l'impératif pluriel alourdit l'énoncé. « Rule of thumbs » — formule idiomatique fautive dans la VO — est rendue par un « règle générale » irréprochable : le système corrige l'approximation de Collins au lieu de la restituer, et efface ainsi une trace d'oralité spontanée, conformément à ce que Chaume décrit comme l'effacement des marqueurs d'oralité préfabriquée (2012, p. 82). « Late game », marqueur sociolectal du jargon gaming, devient « fin de partie » : équivalent descriptif juste, mais dépourvu de la coloration communautaire qui rend le discours crédible auprès de son audience — au même titre que « cringe » rendu « gênant » dans l'extrait 1.
Traduction automatique	Hormis les lacunes de transcription signalées et les glissements de registre mentionnés, la traduction est globalement correcte sur le plan sémantique. La structure informative des conseils donnés par Collins est préservée. Aucune erreur ASR produisant un contresens n'est détectable sur les segments effectivement transcrits.
Relation locuteur–spectateur	La fonction pédagogique de cet extrait — Collins s'adresse directement à son audience pour lui expliquer une mécanique du jeu — est globalement transmise, ce qui constitue une réussite relative par rapport aux extraits précédents. L'adresse directe au spectateur (« As you can see », « Be aware ») est restituée en français sans perte notable de la relation d'adresse. C'est le contexte le plus favorable au doublage automatique dans l'ensemble du corpus analysé jusqu'ici : discours planifié, locuteur unique, syntaxe simple, registre explicatif.
Bilan de l'extrait	
Cas relativement réussi, et analytiquement utile à ce titre : il établit que le dispositif fonctionne mieux sur le discours planifié que sur le discours spontané, conformément au continuum de Chaume (2012, p. 88, Figure 5.1). Les défaillances qui subsistent malgré ces conditions favorables n'en sont que plus significatives. Réussite : fidélité globale, adresse au spectateur préservée. Lacunes : deux pauses de non-transcription, normalisation de « rule of thumbs », neutralisation du jargon (« late game »), glissement de registre sur « abort ».	

FICHE D'ANALYSE — EXTRAIT 10	
Vidéo (titre + chaîne)	Outer Wilds Starter Guide (NO SPOILERS) — Captain Collins

Genre	Gaming
Horodatage de l'extrait	3:49 – 4:12
Durée de l'extrait	23 s
Numéro d'extrait	10
Transcription	
VO (anglais)	That's the computer, yeah go to the computer. That was yesterday. Timber hearth. As you can see right now, let me comeback (rires). As you can see right now there was a map of all the planets you have in the system. So you have to go to every planet that you get to know them. And the other thing that we're now looking at is the rumor mode. And in the rumor mode you have got logs, like for example this one.
Doublage automatique (FR)	C'est l'ordinateur. Ouais, aller à l'ordinateur. C'était hier. Timber. Et comme vous pouvez le voir maintenant, laissez-moi revenir. Comme vous pouvez le voir maintenant, il y avait une carte de toutes les planètes que vous avez dans le système, donc vous devez aller sur chaque planète pour apprendre à les connaître. Et l'autre chose que nous examinons en ce moment est le mode rumeur. Et dans le mode rumeur vous avez des verrous comme par exemple celui-ci.
Observations par critère	
Synchronisation	L'isochronie est globalement maintenue sur cet extrait — aucun blanc notable ne vient interrompre le flux du doublage. En revanche, la superposition de deux couches audio — la voix de Collins en commentaire et l'extrait de stream qu'il diffuse simultanément — pose un problème de segmentation que le système gère de façon approximative. Les répliques issues du stream enchâssé (« That was yesterday. Timber hearth. ») sont traitées par le module ASR comme si elles faisaient partie du discours continu de Collins, sans distinction entre les deux locuteurs ni entre les deux niveaux énonciatifs. Le résultat est une transcription qui fusionne commentaire et stream en un flux indifférencié.
Prosodie & expressivité	L'extrait relève d'un registre explicatif proche de celui de l'extrait 3, ce qui limite les défaillances prosodiques les plus spectaculaires. Le débit est régulier, les variations intonatives modestes. Le rire de Collins lors de « let me comeback » — qui signale qu'il s'amuse de sa propre maladresse — est ignoré par le doublage, effaçant le moment de légèreté dans la continuité du discours explicatif. C'est la même défaillance documentée dans l'extrait 2: les éléments para-verbaux ne sont jamais restitués.
Crédibilité discursive	Cet extrait introduit un phénomène encore non documenté dans l'analyse : la confusion entre deux niveaux énonciatifs distincts. Collins structure sa vidéo en alternant entre son propre commentaire explicatif et des extraits de son stream, qu'il projette à l'écran pour illustrer ses conseils. Pour le spectateur visuel, la distinction est claire — les deux niveaux sont séparés par le montage. Pour le module ASR, qui ne traite que le flux audio sans accès à l'image, cette distinction n'existe pas. Le résultat est une transcription qui fusionne les deux niveaux en un seul discours continu, comme si Collins énonçait « C'était hier. Timber. » dans le cadre de son commentaire pédagogique — ce qui est incompréhensible pour un spectateur francophone qui ne connaît pas le jeu.
Traduction automatique	Deux erreurs méritent une analyse distincte. Premièrement, « Timber Hearth » — nom de la planète de départ dans Outer Wilds, prononcé en deux mots distincts dans la VO — est partiellement transcrit : seul « Timber » est restitué, « Hearth » disparaissant vraisemblablement du fait d'une mauvaise segmentation ou d'une non-reconnaissance du second mot. Le nom propre est ainsi tronqué, perdant la seconde moitié qui le distingue d'un nom commun ordinaire. Pour un

	spectateur francophone découvrant le jeu grâce à ce tutoriel, « Timber » ne renvoie à aucun référent identifiable — il flotte dans le discours sans ancrage. C’est un mécanisme analogue à celui documenté sur « Margit the Fell » dans l’extrait 5 de Maya : le système ne dispose d’aucune connaissance de la terminologie interne du jeu et traite les noms propres comme des séquences sonores ordinaires. Deuxièmement, « logs » est transcrit et traduit par « verrous » — équivalent français de « locks ». Il s’agit d’une erreur de substitution phonétique au niveau de l’ASR : « logs » et « locks » sont suffisamment proches acoustiquement, notamment dans un débit soutenu, pour que le module choisisse le mauvais terme. Conformément au mécanisme décrit par Ruiz et Federico (2014, p. 271), l’erreur de substitution est ensuite transmise telle quelle au module de traduction, qui produit « verrous » — traduction correcte de « locks », mais qui n’a aucun sens dans le contexte d’un système de journaux de notes dans un jeu vidéo. La traduction juste aurait été « journaux » ou « notes ». Le résultat est un contresens complet : un spectateur francophone comprend que le mode rumeur contient des serrures, là où Collins explique qu’il contient des entrées de journal permettant de suivre la progression de l’enquête. Sur les autres répliques, la traduction est sémantiquement correcte. « Let me comeback » est rendu par « laissez-moi revenir » — formulation fonctionnellement juste mais légèrement formelle dans sa conjugaison (« laissez » suggère une adresse à plusieurs personnes de façon distante), là où Collins s’adresse à son audience sur un ton décontracté.
Relation locuteur–spectateur	La fonction pédagogique de cet extrait est partiellement transmise — Collins explique comment fonctionne la carte planétaire et le mode rumeur, et la structure de ses conseils est globalement restituée. Mais l’erreur sur « logs » / « verrous » compromet précisément le passage où Collins introduit un élément central du jeu pour son audience. Le mode rumeur est le système de suivi de la progression narrative d’Outer Wilds — le présenter comme un système de « verrous » produit non seulement un contresens, mais une image du jeu inexacte, susceptible de dérouter un spectateur francophone qui souhaiterait jouer à partir de ce tutoriel.
Bilan de l’extrait	
Deux phénomènes distincts. Le module ASR traite le flux comme un discours continu et ne distingue pas le commentaire de Collins du stream qu’il diffuse simultanément : il fusionne deux niveaux énonciatifs. S’y ajoute une substitution phonétique (« logs » rendu « locks ») qui transforme, dans un tutoriel, l’acte pédagogique en source de désinformation. Réussite : structure informative maintenue, isochronie correcte. Lacunes : fusion des niveaux énonciatifs, troncature de « Timber Hearth », substitution « logs » → « verrous », rire ignoré.	

FICHE D’ANALYSE — EXTRAIT 11	
Vidéo (titre + chaîne)	Outer Wilds Starter Guide (NO SPOILERS) — Captain Collins
Genre	Gaming
Horodatage de l’extrait	6:42 – 6:51
Durée de l’extrait	9 s
Numéro d’extrait	11
Transcription	

VO (anglais)	-And this is all the hint you get. You need... You need to have a look out there. [I call this precise landing. Okay (rire). Extremely precise landing.]
Doublage automatique (FR)	Et c'est tout l'indice pour un atterrissage précis. D'accord, vous devez y jeter un œil. Un (.) atterrissage extrêmement précis.
Observations par critère	
Synchronisation	L'isochronie est approximativement maintenue sur la durée de l'extrait. Le rire étouffé de Collins, audible entre les deux parties de sa réplique dans le stream, est entièrement ignoré, laissant un vide expressif à l'écran sans équivalent sonore dans le doublage.
Prosodie & expressivité	La séquence est brève, construite sur un contraste entre deux registres : le commentaire pédagogique de Collins (posé, didactique) et sa réaction dans le stream (autodérision, rire étouffé, ton légèrement ironique sur « extremely precise landing »). Ce contraste, qui constitue le ressort humoristique du passage, est entièrement absent du doublage. La voix synthétique déroule les énoncés fusionnés sur un débit régulier, ne marquant aucune différence entre le conseil donné à l'audience et la réplique comique adressée à lui-même dans le stream.
Crédibilité discursive	La répétition « You need... You need to have a look out there » est une répétition caractéristique de l'oralité préfabriquée, Collins cherche ses mots, hésite, recommence. Chaume (2012, p. 90) classe précisément ce type de faux départ parmi les marqueurs syntaxiques qui construisent l'authenticité du discours oral dans les productions audiovisuelles. Le doublage efface entièrement cette hésitation, restituant un « vous devez y jeter un œil » lisse et direct, qui ne porte aucune trace de la spontanéité de l'original. Le « Un » isolé en tête de la dernière réplique est un écart à noter également. Le doublage automatique n'a pas fait de liaison entre « Un » et « atterrissage extrêmement précis », ce qui rend l'énoncé robotique lors du doublage.
Traduction automatique	La fusion des deux couches énonciatives, déjà documentée dans l'extrait 4, atteint ici son degré le plus problématique. « And this is all the hint you get » et « I call this precise landing » appartiennent à deux niveaux de discours distincts — le premier est le commentaire de Collins, le second une réplique prononcée pendant son stream et diffusée dans la vidéo à titre d'illustration. Le module ASR les fusionne en un seul énoncé continu, et le module de traduction produit « Et c'est tout l'indice pour un atterrissage précis » — une phrase qui amalgame les deux sources de façon à rendre le propos de Collins partiellement incohérent. Le « hint » (indice) se retrouve ainsi directement associé à l'atterrissage précis, alors que dans la VO, Collins parle de l'indice donné par le jeu sur la façon d'explorer, sans lien direct avec la scène d'atterrissage visible dans le stream.
Relation locuteur–spectateur	L'autodérision de Collins — « I call this precise landing » suivi d'un rire étouffé — est une stratégie relationnelle identique à celle documentée dans l'extrait 2: le créateur rit de sa propre maladresse pour inviter son audience à partager le moment. En ignorant le rire et en fusionnant cette réplique avec le commentaire pédagogique, le doublage neutralise à la fois la connivence et le contraste de registre qui en font l'efficacité. Le spectateur francophone ne perçoit pas qu'il s'agit d'une réplique humoristique — il reçoit une phrase mal construite sans contexte.
Bilan de l'extrait	

Extrait ciblé sur une rupture prosodique interne au syntagme nominal : le doublage restitue « Un (.) atterrissage extrêmement précis », énoncé dont la forme écrite est correcte mais dont la réalisation intercale une pause entre le déterminant et le nom, sans réaliser la liaison. L'anomalie est phonétique, non syntaxique : elle n'empêche pas de comprendre, mais signale une voix qui ne segmente pas le discours selon les unités de sens du français. Réussite : isochronie approximativement maintenue. Lacunes : rupture prosodique et liaison non réalisée, fusion des niveaux énonciatifs, perte du contexte humoristique.

D = Dantes (Doaenel)

TF = Tobias Fate

X = xQc

K = Kadeem

T = Tyler1

FICHE D'ANALYSE — EXTRAIT 12	
Vidéo (titre + chaîne)	xQc Lost his MIND after Kadeem said this to him... ft Tyler1 & Tobias Fate — Doaenel
Genre	Gaming
Horodatage de l'extrait	22:20 – 22:44
Durée de l'extrait	24 s
Numéro d'extrait	12
Transcription	
VO (anglais)	TF- Rakan mid, top. D- [No, but] TF- CC, AD Carry. [I mean] D- I mean the the only thing that it works against. So the only thing is uh Xayah's root. It doesn't do anything versus Rakan's W because he just chains both anyways, and like everything else is either a snare or a knock up. TF- [This guy is fucking (xxx) god damn!] X- Jesus Christ this guy. I mean (xxx). D- [BRO, YOU'RE FUCKING ASKING]. Okay, Tyler, play up, play up up up. T- [I have (xxx) up.]
Doublage automatique (FR)	Raon mid, top. Non, mais c'est CD Carry. Je veux dire, la seule chose contre laquelle ça marche, donc la seule chose c'est ZiOUS. Ça ne fait rien contre le W de reconnaissance, parce qu'il enchaîne les deux de toute façon et tout le reste. C'est soit un S soit un knock-up. De Jésus-Christ ce mec. Je veux dire, -[BRO] -c'est ce que - [BRO YOU'RE FUCKING] tu demandes. D'accord.

Observations par critère	
Synchronisation	Cet extrait pousse à ses limites extrêmes les défaillances de synchronisation déjà documentées dans les analyses précédentes. La multiplication des locuteurs — cinq hommes en vocal simultané — et la densité des chevauchements rendent l'isochronie structurellement inaccessible pour un système conçu pour traiter un flux audio monologique. Le module ASR ne segmente pas les tours de parole : il produit un flux transcrit continu qui fusionne les interventions de Tobias Fate, Dantes, xQc et Tyler1 en un seul énoncé attribué à une voix unique. Le cas le plus révélateur est celui du cri de Dantes (« BRO, YOU'RE FUCKING ASKING »). Son volume dans la bande audio est manifestement trop élevé pour que le module de traitement le supprime au profit de la voix synthétique. Le résultat est que les fragments originaux en anglais — « BRO » et « BRO YOU'RE FUCKING » — restent audibles dans la version française, intercalés entre les segments doublés. Le spectateur entend simultanément la voix doublée et la voix originale, sans que l'une n'efface l'autre, exposant directement au spectateur le fonctionnement et les limites du dispositif. Ce phénomène est qualitativement différent des cas documentés sur les vidéos précédentes, où la VO résiduelle constituait un fond atténué : ici, l'original surgit au premier plan et coexiste avec le doublage dans le même espace sonore.
Prosodie & expressivité	La dimension prosodique de cet extrait est anéantie à deux niveaux. D'une part, l'attribution d'une voix féminine unique à cinq locuteurs masculins — déjà signalée comme défaillance systémique pour l'ensemble de la vidéo — supprime toute différenciation entre les locuteurs. Il est impossible pour le spectateur francophone de reconstituer qui parle à quel moment, même lorsque les interventions se succèdent sans chevauchement. D'autre part, le contraste entre le débit analytique et posé de Dantes (qui explique une mécanique du jeu) et les exclamations de Tobias Fate, xQc et Tyler1 (qui réagissent à l'action en temps réel) est intégralement effacé. La voix synthétique déroule l'ensemble sur un débit uniforme et sans inflexion, transformant un échange chaotique et expressif en une succession d'énoncés intonativement indistincts. Le cri de Dantes — « BRO, YOU'RE FUCKING ASKING » — est l'un des moments d'intensité émotionnelle les plus marqués de l'extrait. Sa force dans la VO tient à la conjonction de plusieurs facteurs : volume, débit accéléré, registre familier voire agressif, adresse directe à Tyler1. Dans le doublage, ce moment n'est pas restitué — il émerge dans son format original en anglais, encadré par des fragments de la voix synthétique. L'expressivité du cri n'est donc ni traduite ni incarnée : elle est simplement laissée telle quelle, dans une langue que le doublage était précisément censé remplacer.
Crédibilité discursive	Le discours de Dantes est dense en termes propres à League of Legends — « CC » (crowd control, capacités qui entravent les déplacements adverses), « AD Carry » (rôle de tireur), « Xayah's root » (capacité d'immobilisation), « Rakan's W » (deuxième capacité du champion), « knock up » (projection en l'air) —, immédiatement identifiables pour un joueur et opaques pour tout autre. Aucun n'est correctement restitué : le détail des substitutions figure dans la case consacrée à la traduction automatique. Ce qui importe ici est l'effet produit. Le référentiel commun sur lequel repose l'échange entre Dantes et son audience disparaît intégralement, et le spectateur francophone se trouve devant un discours dont il ne peut reconstituer ni l'objet ni la logique. La crédibilité du locuteur, qui tient à sa maîtrise reconnue de ce vocabulaire, s'effondre avec lui.
Traduction automatique	Cet extrait constitue le cas le plus documenté d'effondrement de la chaîne ASR-TAN dans l'ensemble du corpus. Le mécanisme décrit par Ruiz et Federico (2014, p. 266) — propagation des erreurs de transcription sans correction possible aux étapes suivantes — atteint ici un degré où plusieurs erreurs s'accumulent sur un seul énoncé, rendant impossible toute reconstruction du sens par le spectateur. Deux facteurs aggravent ce mécanisme dans ce contexte spécifique. D'une part, on peut supposer que le jargon de League of Legends est quasi absent des

	corpus d'entraînement des systèmes de TAN standard : des termes comme « root », « knock up », « CC » ou des noms de champions comme « Xayah » et « Rakan » auraient une distribution extrêmement restreinte dans les textes écrits sur lesquels ces systèmes sont entraînés. D'autre part, les chevauchements de voix dégradent la qualité du signal audio reçu par le module ASR, augmentant mécaniquement le taux d'erreur de reconnaissance sur l'ensemble de l'extrait — les erreurs de substitution ainsi produites étant, selon Ruiz et Federico (2014, p. 271), particulièrement coûteuses pour la qualité de la traduction finale. Le détail des substitutions est le suivant : « CC » (crowd control) est rendu « CD », sigle qui désigne en réalité le cooldown, de sorte que la fonction du terme est remplacée par une notion différente ; « AD Carry » n'est maintenu que partiellement et se trouve précédé de ce « CD » erroné, ce qui rend le syntagme incohérent ; « Xayah's root » devient « Ziou », forme qui n'existe ni dans le jeu ni en français ; « Rakan's W » devient « W de reconnaissance » ; et « snare » est réduit à la lettre « S ».
Relation locuteur–spectateur	L'uniformisation vocale féminine pour cinq locuteurs masculins est la défaillance relationnelle la plus radicale documentée dans l'ensemble du corpus. Elle dépasse les cas d'attribution incorrecte de genre observés sur la vidéo de Maya, où une voix masculine était attribuée à une locutrice féminine : ici, c'est l'ensemble des identités vocales de cinq personnalités distinctes — xQc, Tyler1, Tobias Fate, Kadeem et Dantes — qui sont effacées au profit d'une voix unique. Or ces streamers sont des figures dont la voix constitue un marqueur identitaire fort pour leur audience respective. xQc est reconnaissable à son accent québécois et à son débit très rapide ; Tyler1 à son élan d'énervement intenses et à sa manière de parler. Pour un spectateur familier de ces créateurs, l'unification vocale est perçue immédiatement comme une rupture radicale. Pour un spectateur non familier, elle rend impossible toute identification des locuteurs.
Bilan de l'extrait	
Effondrement complet du dispositif en conditions de complexité maximale : cinq locuteurs simultanés, chevauchements constants, jargon hyperspécialisé. Aucun critère de la grille n'est satisfait. Le résultat n'est plus un doublage dégradé mais un artefact sonore qui ne remplit aucune des fonctions du discours original — ni informative, ni expressive, ni relationnelle. Pérez-Ortiz et al. (2022, p. 146) posent que la qualité d'un système de TAN se mesure à son adéquation à la fonction communicative du texte source : elle est ici nulle. Aucune réussite documentable. Lacunes : voix féminine unique pour cinq locuteurs masculins, cinq erreurs terminologiques majeures, fusion des tours de parole, cri de Dantes laissé en anglais.	

FICHE D'ANALYSE — EXTRAIT 13	
Vidéo (titre + chaîne)	xQc Lost his MIND after Kadeem said this to him... ft Tyler1 & Tobias Fate — Doael
Genre	Gaming
Horodatage de l'extrait	29:45 – 30:05
Durée de l'extrait	20 s
Numéro d'extrait	13
Transcription	
VO (anglais)	D- Lowkey Kadeem, get Zhonya's. Zhonya's is actually a good call. TF- [I'm fucking this fucker, I don't care.] D- Go Zhonya's, [go Zhonya's]

	TF- Nah, don't bo- this guy's a wank. T- I don't give a shit, I don't care. K- No, n° (xxx) T- Wait, wait, wait, wait dragon's good actually. X- [Why you're just trolling him with these dumbass ideas?] K- (xxx) and botlane so like if I do armor, you guys won the match. D- No, it's not trolling X, at this point, if he goes Zhonya's T- (xxx) backup. D- it's actually really good, because if they try to engage him, he Zhonya's. xQc- (xxx)
Doublage automatique (FR)	Lowkey pourrait même avoir Zas. Zombie est en fait un bon choix. Vas-y dans cette zone. Ce type est un branleur, je m'en fous, je n'attends (...) pas de le troller avec ces (débit accéléré) d'armure et vous avez gagné le match. Non, ce n'est pas le cas, il ne troll pas X à ce stage. S'il y va, c'est vraiment bien, ils essaient. Eux ouais.
Observations par critère	
Synchronisation	Sur vingt secondes d'extrait, les chevauchements sont quasi permanents. Le module ASR, incapable de segmenter des tours de parole simultanés, produit une transcription continue qui sélectionne arbitrairement certains segments et en ignore d'autres, sans logique identifiable. La pause (...) dans le doublage signale un moment où le système a renoncé à transcrire — vraisemblablement un passage où la superposition des voix rend le signal audio trop dégradé pour être exploitable. L'accélération soudaine du débit documentée un peu plus loin indique que le système tente de compenser les blancs introduits par ses propres lacunes de transcription, selon le même mécanisme observé dans l'extrait 1 de Captain Collins — mais ici dans un contexte bien plus chaotique, ce qui rend la correction d'autant plus mécanique et perceptible.
Prosodie & expressivité	L'extrait concentre plusieurs registres émotionnels simultanés que le doublage écrase uniformément. Dantes se montre analytique et persuasif, il tente de convaincre les autres que Kadeem doit acheter un Zhonya's. Tobias Fate et Tyler1 sont concentrés et même agacés. xQc intervient de manière ironique et lance une pique à Dantes. Kadeem tente répondre. Ces cinq colorations émotionnelles distinctes sont absentes du doublage, restitué sur une voix féminine uniforme et sans inflexion, au débit régulier sauf lors de l'accélération algorithmique. La dynamique de l'échange est entièrement neutralisée.
Crédibilité discursive	L'extrait introduit une défaillance inédite : le même terme est transcrit de deux façons différentes à quelques secondes d'intervalle. « Zhonya's », nom d'un objet du jeu prononcé trois fois par Dantes en vingt secondes, devient « Zas » puis « Zombie ». Le système ne dispose donc pas seulement d'une mémoire terminologique déficiente, déjà documentée à l'extrait précédent : il n'assure aucune cohérence interne d'une phrase à l'autre. Pour le spectateur, le même référent change de nom en cours d'échange, ce qui rend le discours non seulement obscur mais incohérent avec lui-même. Les autres termes spécialisés subissent un sort comparable, « botlane » disparaissant entièrement.
Traduction automatique	L'effondrement de la chaîne ASR-TAN est ici aggravé par un facteur documenté pour la première fois dans l'analyse : la variabilité interne de la transcription sur un même terme. « Zhonya's » devient « Zas » puis « Zombie » — deux erreurs distinctes sur le même mot. Cette incohérence interne suggère que le module ASR traite chaque segment audio de façon véritablement indépendante, sans mémoire de ses propres transcriptions antérieures dans la même séquence. Pour un spectateur francophone, le résultat est une impossibilité supplémentaire : même si l'un des deux termes produits pouvait faire sens dans un contexte donné, leur contradiction interne dans l'espace de quelques secondes rend tout décodage impossible.

	La fragmentation du discours de Dantes illustre par ailleurs une limite complémentaire : son argument, qui s'étend sur plusieurs tours de parole entrecoupés d'interventions des autres joueurs, est transcrit de façon morcelée. Les segments retenus (« Non, ce n'est pas le cas, il ne troll pas X à ce stage. S'il y va, c'est vraiment bien, ils essaient. ») sont syntaxiquement cohérents pris isolément, mais privés de leur contexte — notamment la mention de Zhonya's, rendue incompréhensible — ils ne permettent pas de reconstituer le raisonnement de Dantes.
Relation locuteur–spectateur	L'uniformisation vocale féminine documentée comme défaillance systémique de cette vidéo atteint ici ses conséquences les plus concrètes. L'échange entre les cinq joueurs est un débat stratégique où chaque voix représente une position distincte : Dantes défend Zhonya's, Tyler1 et Tobias Fate s'y opposent ou s'en désintéressent, Kadeem apporte un argument sur la botlane, xQc ironise. Pour l'audience des streamers, suivre cet échange, c'est suivre cinq personnalités dont les opinions divergent — et la dynamique entre ces personnalités est précisément ce que le contenu vend. En réduisant les cinq voix à une seule, le doublage ne produit pas seulement une erreur d'attribution : il détruit la structure argumentative de l'échange, le transformant en un monologue incohérent dont le locuteur change sans que rien ne le signale.
Bilan de l'extrait	
Extrait ciblé : le même terme (Zhonya's) est transcrit de deux façons différentes à quelques secondes d'intervalle. Le système ne dispose d'aucune mémoire terminologique interne et traite chaque segment comme un événement sonore isolé. Réussite : « wank » rendu « branleur », équivalent de registre correct. Lacunes : deux transcriptions divergentes du même terme, fusion des cinq locuteurs en une voix unique, fragmentation de l'argument, disparition de plusieurs passages.	

FICHE D'ANALYSE — EXTRAIT 14	
Vidéo (titre + chaîne)	xQc Lost his MIND after Kadeem said this to him... ft Tyler1 & Tobias Fate — Doael
Genre	Gaming
Horodatage de l'extrait	1:03:02 – 1:03:41
Durée de l'extrait	39 s
Numéro d'extrait	14
Transcription	
VO (anglais)	TF- He presses the alt button and becomes a tank. X- You wanna smoke this guys or not? TF- I don't do damage, I wanna play with you, but we have to go get someone else to do damage. X- Yeah D- Also, Tyler- K- I'm gonna try to get this guy D- Tyler, don't forget- X- Jax is gonna be top, he's gonna be top 100%, let's go top. D- Also, Tyler, do not forget, we need you to get Cleaver next item, I need you to get cleaver for me. T- I'm not on cleaver bro. K- You want me to go Cleaver?

	D- Yeah Kadeem, you can actually get it too. T- (xxx) no man. D- Guys, we should look to play on, we should (xxx) TF- [We gotta do something] about Kassadin, like I would never be able to beat Kassadin. T- Do you want me to go match Kassadin, then you run around? TF- Well, no, no, no, I mean, me and Dantes have to kill this guy. We have to kill him. X- [(xxx)I think we lose one v three against (xxx)] T- What do you want from me? (singing)
Doublage automatique (FR)	Il appuie sur le bouton alt et devient un tank. Tu veux fumer ce type ? je ne fais pas de dégats, je veux jouer avec toi, mais il faut qu'on aille chercher quelqu'un d'autre pour faire des dégats. On ne peut pas tuer. Je vais essayer d'obtenir ce sommet. Jack va aller en haut, il va réussir un eagle à 100%. Allons en haut, ouais. N'oubliez pas non plus qu'il nous faut le prochain objet Cleaver. Il me faut Cleaver. Je ne suis pas malin, ouais tu peux l'obtenir aussi. Je ne sais pas, mec. Les gars, on devrait essayer de jouer, on devrait faire quelque chose avec Kassadin, genre je ne pourrais jamais battre Kassadin. D'accord, tu veux aller affronter Kassadin et courrir partout ? Eh bien, non, non, non, je veux dire Dante et moi devons tuer ce type, on doit le tuer, je n'ai même pas perdu un seul v trois contre. Qu'est-ce tu veux de moi ?
Observations par critère	
Synchronisation	Sur près de quarante secondes d'échange multi-locuteurs, le doublage produit un flux relativement continu — mais au prix d'une fusion systématique des tours de parole qui rend impossible toute identification des locuteurs. Plusieurs interventions sont omises ou tronquées, notamment les passages signalés par (xxx). Le morcellement de l'échange original est masqué par la continuité artificielle du flux doublé, qui enchaîne les fragments retenus sans marquer les ruptures ni les chevauchements.
Prosodie & expressivité	La séquence concentre des registres variés : la tentative de coordination stratégique de Dantes, l'exaspération de Tobias Fate face à la situation en jeu, l'ironie de xQc, et l'humour involontaire de Tyler1 qui finit par chanter sa réplique. Ce dernier moment — « What do you want from me ? (singing) » — est l'un des marqueurs expressifs les plus singuliers de l'extrait. Sa restitution par un simple « Qu'est-ce tu veux de moi ? » en voix synthétique neutre efface entièrement la dimension comique et relationnelle de la réplique. La voix de synthèse ne dispose d'aucun mécanisme pour modéliser le passage du discours au chant — confirmant ce que Schröder (2009, p. 111) identifie comme limite structurelle des voix de synthèse face aux variations pragmatiques et expressives de la parole humaine.
Crédibilité discursive	L'écart le plus significatif tient à la confusion entre « cleaver » et « clever ». Tyler1 dit « I'm not on cleaver bro », signifiant qu'il n'achètera pas l'objet Black Cleaver ; le module ASR transcrit « clever », et la traduction produit « je ne suis pas malin ». La traduction est correcte au regard de ce qu'elle reçoit — c'est la transcription qui est fautive. Le résultat est un énoncé grammaticalement et pragmatiquement normal, mais sans rapport avec l'original. Là où « Ziou » ou « le marché qui tombe » signalaient immédiatement leur caractère aberrant, celui-ci passe le filtre de la vraisemblance : un spectateur ignorant la version originale ne peut pas le détecter. S'y ajoute une double inconsistance sur les noms propres — « Kassadin » correct à la première occurrence puis « Kassidin », « Jax » rendu « Jack » —, ainsi que « go top » rendu littéralement « allons en haut », qui efface la dimension stratégique du terme, et un « eagle » surgi d'aucun segment identifiable de la VO.
Traduction automatique	L'erreur « cleaver » → « clever » → « je ne suis pas malin » est analytiquement la plus riche de l'ensemble du corpus, précisément parce

	qu'elle illustre un mécanisme distinct de ceux déjà documentés. Ruiz et Federico (2014, p. 271) distinguent les erreurs de substitution selon leur nature : les substitutions par des mots grammaticaux sont les plus coûteuses pour la traduction finale, mais les substitutions par des mots lexicaux phonétiquement proches — comme ici — sont particulièrement difficiles à détecter, car elles produisent des résultats superficiellement cohérents. Le système ne peut pas identifier l'erreur : il reçoit « I'm not clever » et traduit correctement.
Relation locuteur–spectateur	La réplique finale de Tyler1 — « What do you want from me? (singing) » — mérite une attention particulière sur ce critère. Le passage au chant est une stratégie relationnelle caractéristique du personnage de Tyler1, dont l'humour repose souvent sur des ruptures de registre inattendues. Pour son audience, cette réplique est immédiatement reconnaissable comme une signature de sa personnalité. En la restituant sur une voix synthétique neutre, le doublage n'efface pas seulement un élément para-verbal : il supprime un marqueur identitaire, renforçant l'effacement des cinq personnalités vocales distinctes au profit de la voix féminine uniforme documentée depuis le début de l'analyse de cette vidéo.
Bilan de l'extrait	
L'erreur « cleaver » transcrite « clever » puis rendue « je ne suis pas malin » est la plus révélatrice du corpus. Contrairement aux erreurs aberrantes, immédiatement identifiables comme telles, celle-ci produit un énoncé superficiellement cohérent qu'un spectateur ignorant la version originale ne peut pas détecter. Elle établit que la vérification humaine est nécessaire à chaque étape de la chaîne, et non seulement en cas de résultat visiblement fautif. Réussite : « Kassadin » transcrit correctement une fois, « Cleaver » maintenu deux fois. Lacunes : substitution homophonique produisant un contresens indétectable, inconsistances sur les noms propres, réplique chantée ignorée, uniformisation vocale maintenue.	

8.3 Synthèse du doublage automatique face aux vidéos gaming

L'analyse des quatorze extraits du sous-corpus gaming fait apparaître un ensemble de constantes dont la récurrence, au-delà des différences entre vidéos, révèle des limites structurelles plutôt que des défaillances ponctuelles.

La neutralisation prosodique est la défaillance la plus systématique de l'ensemble du corpus. Sur la totalité des extraits, la voix de synthèse restitue le contenu verbal des énoncés sans en reproduire la coloration émotionnelle ni le contour intonatif. Les rires, les montées expressives, les décélérations rhétoriques et les accélérations spontanées, qui portent une large part du sens dans un contenu gaming, sont soit ignorés, soit aplatis en un débit régulier. L'extrait 8, analysé dans un contexte volontairement favorable — locuteur unique, registre léger, syntaxe simple —, le montre sans ambiguïté : même lorsque la transcription et la traduction sont correctes, la voix synthétique n'incarne pas la réaction spontanée. Ce que Schröder (2009, p.

111) identifie comme limite fondamentale des voix de synthèse se confirme à chaque extrait, indépendamment du type de contenu.

Les erreurs de la chaîne ASR-TAN constituent le second axe de défaillance, et leur nature varie selon la configuration du discours. Sur la vidéo de Maya, elles touchent principalement les noms propres issus de l'univers d'*Elden Ring* : le système ignore les localisations françaises officielles et produit des reconstructions phonétiques approximatives (« Dung Eater » rendu « la perte de l'arme », « Margit the Fell » rendu « le marché qui tombe »). Ces erreurs ne sont pas anodines, car les noms de personnages et les mécaniques de jeu constituent le référentiel commun entre le créateur et son audience. Sur la vidéo de Captain Collins, les erreurs touchent davantage la segmentation du discours : le système, incapable de distinguer le commentaire de Collins des extraits de stream qu'il diffuse simultanément, fusionne deux niveaux énonciatifs en un flux continu et produit des énoncés dont la cohérence référentielle est compromise ; sur le même extrait, la voix de synthèse intercale en outre une pause entre le déterminant et le nom, sans réaliser la liaison attendue (« Un (.) atterrissage extrêmement précis »). Sur la vidéo de Doael, l'accumulation de chevauchements et de jargon hyperspécialisé pousse le module ASR à ses limites : le même terme (Zhonya's) est transcrit de deux façons différentes dans le même extrait, et certaines substitutions phonétiques produisent des énoncés superficiellement cohérents mais sémantiquement faux (« cleaver » transcrit « clever », rendu « je ne suis pas malin »). Ce dernier type d'erreur est analytiquement le plus insidieux : là où la majorité des erreurs signalent immédiatement leur caractère aberrant, celles-ci franchissent le filtre de la cohérence de surface, ce qui rend leur détection impossible sans accès à la version originale.

L'incapacité du système à maintenir une identité vocale cohérente entre les locuteurs constitue la troisième défaillance majeure. Sur l'ensemble du corpus, le doublage automatique ne distingue pas les locuteurs de façon stable : une voix masculine est attribuée à des locutrices (extraits 1 à 3 de Maya), une voix féminine unique est utilisée pour cinq locuteurs masculins (Doael), et un changement de voix survient à l'intérieur d'une même réplique (extrait 6 de Maya). Cette instabilité révèle l'absence de tout module d'identification des locuteurs dans la chaîne de traitement : le système attribue une voix de synthèse disponible à l'ensemble du flux audio, sans tenir compte des corps visibles à l'écran ni des identités vocales reconnaissables pour l'audience. Elle contrevient à un standard formulé par Díaz Cintas et rapporté par Chaume, selon lequel les mêmes voix ne doivent pas être réutilisées à travers une production ni d'une production à l'autre (Díaz Cintas 2003, pp. 55-58, cité dans Chaume 2012, p. 20). Chaume

souligne par ailleurs que la réussite d'un produit doublé dépend du choix des voix appropriées, décision que le directeur de plateau prend au cas par cas (2012, p. 36, cité dans Bosseaux 2018, p. 52) : le doublage automatique ne comporte aucune étape équivalente. Dans des contenus gaming où la relation entre un créateur et ses spectateurs repose sur la reconnaissance de sa voix comme signature identitaire, cette uniformisation est particulièrement coûteuse. Comme le formule Bosseaux, remplacer une voix par une autre revient à déconstruire le lien entre le corps et la voix (2015, p. 110) — condition que le doublage automatique ne remplit à aucun moment sur ce corpus.

Un quatrième constat traverse l'ensemble des analyses : la voix originale reste partiellement audible en fond sonore sur la majorité des extraits. Le système d'atténuation du signal source ne supprime pas la voix originale, il la réduit à un fond atténué qui coexiste avec la voix synthétique. Sur les moments d'intensité vocale élevée (cri de Dantes dans l'extrait 12, montée dans les aigus de Collins dans l'extrait 8), le signal original perce le filtre d'atténuation et devient directement perceptible au-dessus du doublage. Le spectateur entend alors simultanément deux couches sonores dans deux langues différentes. Cette caractéristique évoque la logique du voice-over, où la voix originale est traditionnellement maintenue à un volume réduit (Díaz Cintas et Orero 2010, p. 441), mais elle s'en distingue par l'absence de l'organisation réglée qui régit ce mode : ni attaque en clair, ni reprise du volume en fin d'énoncé. On y reviendra au chapitre 11.

Ces quatre constantes — neutralisation prosodique, erreurs en cascade de la chaîne ASR-TAN, instabilité vocale et coexistence des deux bandes sonores — sont présentes sur l'ensemble du corpus, mais leur degré varie significativement selon la configuration communicationnelle de chaque vidéo. La vidéo de Captain Collins, avec un locuteur unique et un registre planifié, fournit les résultats les moins dégradés : les segments explicatifs simples sont correctement transcrits et traduits, et le doublage remplit partiellement sa fonction informative. La vidéo de Maya introduit la pluralité de locuteurs, le jargon spécialisé et les réactions spontanées, ce qui multiplie les occasions d'erreur et compromet systématiquement la dimension relationnelle de la communication. La vidéo de Doaenel, avec cinq locuteurs simultanés et un jargon pour lequel le système ne dispose d'aucune ressource terminologique, produit un résultat qui n'est plus un doublage dégradé mais un artefact sonore ne remplissant aucune des fonctions communicatives du discours original. Cette gradation confirme que les défaillances du système ne sont pas

distribuées de façon aléatoire : elles s'aggravent en fonction de deux variables identifiables — la densité de l'expressivité orale et la complexité de la configuration sonore.

Ce que l'ensemble du corpus gaming permet finalement de constater, c'est que le doublage automatique de YouTube produit des résultats acceptables uniquement dans les conditions les plus restrictives : un locuteur seul, un registre neutre ou planifié, une syntaxe simple, l'absence de jargon spécialisé et l'absence de chevauchements. Ces conditions correspondent à une fraction réduite des contenus gaming effectivement produits sur YouTube. Dès qu'un seul de ces paramètres est mis en défaut — et dans le gaming, plusieurs le sont simultanément —, le système produit un transfert linguistique dont la fidélité au texte source masque l'écart avec l'acte communicationnel original. Marzà et Chaume rappellent que ce qui importe dans un doublage est l'impression produite et la crédibilité de la parole saisie comme un tout, davantage que la synchronisation formelle (2009, p. 35) : sur ce critère, le doublage automatique de YouTube échoue sur la quasi-totalité des extraits analysés.

9. Analyse du sous-corpus de vulgarisation scientifique

Ce chapitre présente l'analyse des sept extraits issus des deux vidéos du sous-corpus de vulgarisation scientifique, selon la même organisation que le chapitre précédent.

9.1 Présentation des extraits analysés

Les vidéos du sous-corpus de vulgarisation scientifique

Le sous-corpus de vulgarisation scientifique regroupe deux vidéos représentant des configurations communicationnelles distinctes, choisies pour confronter le doublage automatique à un registre soutenu, une terminologie spécialisée et des formats d'adresse différents de ceux du gaming.

World Science Festival – *Straight Talk on Cosmic Origins* (YouTube, 10 janvier 2025 ; 1 h 34 min 54 s ; 123 437 vues ; 1,39 M abonnés)

Cette vidéo est une production YouTube native : une interview en visioconférence entre le physicien et vulgarisateur Brian Greene et le cosmologiste Andreas Albrecht, retransmise telle quelle sur la chaîne World Science Festival. Le format est non édité et le discours semi-préparé — les deux intervenants développent leurs propos de façon spontanée autour de questions formulées à l'avance. L'écran est scindé en deux, maintenant les deux locuteurs visibles simultanément tout au long de la vidéo. Ce dispositif visuel impose au doublage automatique une contrainte d'incarnation vocale particulièrement exigeante : le spectateur voit en permanence deux corps distincts auxquels le système doit, idéalement, faire correspondre deux voix distinctes. Le registre est académique et introspectif, avec une oralité spontanée dense en marqueurs para-verbaux — hésitations, rires, dysfluences — qui caractérisent un échange intellectuel en temps réel plutôt qu'un exposé préparé.

Big Think – *Brian Cox on how black holes could unlock the mysteries of our universe* (YouTube, 27 octobre 2023 ; 12 min 51 s ; 4 856 709 vues ; 8,84 M abonnés)

Cette vidéo est un monologue préparé produit par la chaîne Big Think, format caractéristique de la plateforme : un expert s'exprime face caméra, de façon continue, sur un sujet de sa spécialité. Brian Cox, professeur de physique des particules à l'Université de Manchester, présente les trous noirs et les concepts fondamentaux de la cosmologie. Le discours est entièrement planifié, la syntaxe soignée, et le registre oscille entre le soutenu et le

pédagogique. L'écart entre la valeur observée Avec un locuteur unique et un débit maîtrisé, ce format représente les conditions *a priori* les plus favorables du sous-corpus de vulgarisation — l'équivalent de ce que la vidéo de Captain Collins représentait pour le gaming.

Critères de sélection des extraits

Sept extraits ont été retenus pour le sous-corpus de vulgarisation scientifique : quatre pour la vidéo du World Science Festival, trois pour celle de Big Think. La même logique de sélection que pour le sous-corpus gaming est appliquée, combinant des extraits ciblés, qui isolent un phénomène précis, et des extraits transversaux, qui permettent d'observer plusieurs critères en interaction simultanée.

Pour le World Science Festival, la durée de la vidéo (1 h 34 min 54 s) a impliqué un visionnage repérage accéléré, conformément au protocole défini pour les vidéos longues. Les extraits ont été sélectionnés aux moments où le discours spontané est le plus dense — hésitations, rires, reformulations, répliques brèves — et aux passages où la terminologie scientifique ou les références académiques risquaient de poser problème au module ASR. Pour Big Think, la durée plus courte et le format monologique ont permis un visionnage intégral ; les trois extraits retenus couvrent le début de la vidéo (présentation), un passage explicatif à terminologie dense (vitesse de libération, noms propres historiques) et la conclusion rhétorique de la vidéo (citation d'Einstein).

Un seul extrait est ciblé dans ce sous-corpus (extrait 5) : il correspond au passage de Big Think où le doublage produit son résultat le plus réussi de l'ensemble du corpus, ce qui constitue en soi un phénomène analytiquement significatif. Sa sélection permet d'identifier les conditions précises sous lesquelles le système fonctionne de façon acceptable, et de contraster ce résultat avec les défaillances documentées sur les extraits adjacents.

Tableau récapitulatif des extraits

T = transversal (plusieurs critères en interaction) · C = ciblé (phénomène précis) · Pro = prosodie et expressivité · TA = traduction automatique / ASR · Disc = crédibilité discursive · Rel = relation locuteur-spectateur · Syn = synchronisation

N°	Vidéo	Horodatage	Durée	Type	Phénomènes principaux	Critères
1	WSF	2:24 – 2:51	27 s	T	Erreur TTS sur « biochimie » prononcé « biochimia » ; uniformisation vocale masculine pour deux intervenants visibles simultanément ; voix originale d'Albrecht audible en fond	Disc · Rel · Syn
2	WSF	7:55 – 8:35	40 s	T	Intonation déclarative sur réplique de Brian intentionnellement suspendue ; « et ou » produit par transcription erronée d'une hésitation para-verbale ; normalisation des dysfluences	Pro · Disc · TA
3	WSF	10:55 – 11:40	45 s	T	Erreur ASR sur « Phys Rev D » ; désambiguïsation lexicale erronée (« talk » → verbe) produisant « Bob Dickey parle bien » ; chevauchement ignoré	TA · Disc · Pro
4	WSF	15:50 – 16:24	34 s	T	Perçage de la voix originale lors d'un pic d'intonation aiguë ; « euh » synthétique parasitant le discours ; segmentation de la phrase conditionnelle en deux unités distinctes	Pro · Syn · Disc
5	Big Think	1:06 – 1:20	14 s	C	Doublage quasi irréprochable ; non-traduction correcte des entités nommées (Royal Society, titre du livre) ; erreur TTS phonologique sur le titre anglais prononcé à la française	Disc · TA
6	Big Think	1:46 – 2:23	37 s	T	Erreur TTS sur « Laplace » prononcé « Liasse » ; substitution ASR « miles » → « mètres » produisant un ordre de grandeur erroné d'un facteur mille ; scission de la phrase conditionnelle	TA · Disc · Pro
7	Big Think	11:59 – 12:18	19 s	T	Erreur TTS de désambiguïsation homographique (« fils » au sens de fil, prononcé [fis] au lieu de [fil]) ; segmentation de la phrase conditionnelle en deux unités ; neutralisation de la prosodie rhétorique délibérée de Cox	Disc · Pro · TA

Lecture du tableau

La quasi-totalité des extraits sont transversaux, à l'exception de l'extrait 5. Ce déséquilibre entre ciblé et transversal n'est pas le même que pour le gaming — il tient moins à la complexité du discours qu'à la nature des défaillances observées, qui affectent rarement un seul critère de façon isolée. Une erreur TTS sur un nom propre (Laplace → Liasse, extrait 6) entraîne simultanément une perte de crédibilité discursive et une rupture de la relation entre le savant évoqué et son image dans la mémoire du spectateur ; une segmentation erronée de phrase (extraits 2, 6, 7) affecte à la fois la prosodie et la crédibilité du discours. Les défaillances du sous-corpus scientifique sont moins spectaculaires que celles du gaming, mais elles restent rarement isolables sur un critère unique.

Une différence notable avec le sous-corpus gaming tient à la distribution des erreurs entre les trois modules de la chaîne de traitement. Dans le gaming, les erreurs se situaient principalement au niveau de l'ASR, aggravées par la densité terminologique et les chevauchements. Dans ce sous-corpus, des erreurs au niveau du module TTS lui-même apparaissent de façon récurrente — prononciation incorrecte de « biochimie », de « Laplace », de « fils » au sens de fil, et phonologie française appliquée à un titre anglais. Ces erreurs TTS, absentes du sous-corpus gaming, indiquent que la chaîne de traitement est vulnérable à des points distincts selon le type de contenu, et que la vulgarisation scientifique expose

particulièrement les limites du module de synthèse vocale face à la terminologie soutenue et aux noms propres d'origine étrangère.

9.2 Analyse des extraits

A = Andreas Albrecht

B = Brian Greene

FICHE D'ANALYSE — EXTRAIT 1	
Vidéo (titre + chaîne)	Straight Talk on Cosmic Origins — World Science Festival
Genre	Vulgarisation scientifique
Horodatage de l'extrait	2:24 – 2:51
Durée de l'extrait	27 s
Numéro d'extrait	1
Transcription	
VO (anglais)	A- And uh, you know, that was at a time I would imagine there weren't too many women who were teaching biochemistry in those days. B- That's right, and one of the pieces I wish I had asked her when she was around was um what it was like to after to drop for 20 years and then catch up on biochemistry in in 19 (rire) you know 78 or whenever she was doing it.
Doublage automatique (FR)	A- Et vous savez c'était à une époque où j'imagine qu'il n'y avait pas beaucoup de femmes qui enseignaient la biochimie. B- L'époque, c'est vrai. Et l'une des questions que j'aurais aimé lui poser quand elle était là, c'était ce que ça faisait d'abandonner pendant vingt ans et de rattraper son retard en biochimie en 19, vous savez, 78 ou à chaque fois qu'elle le faisait.
Observations par critère	
Synchronisation	L'écran scindé, qui maintient les deux intervenants visibles simultanément tout au long de l'entretien, pose un problème spécifique à ce format. Lorsque Albrecht parle, sa voix originale reste partiellement audible entre les prises du doublage, interférant avec la voix synthétique — phénomène déjà documenté sur d'autres extraits du corpus, mais ici particulièrement perceptible parce que le spectateur voit Albrecht articuler en temps réel. Les lèvres à l'écran et la voix entendue ne coïncident à aucun moment. L'isochronie est globalement maintenue sur les répliques restituées, sans blanc notable.
Prosodie & expressivité	Le registre de cet extrait est celui d'une conversation entre deux académiciens — posée, réflexive, avec quelques marques d'oralité (« uh », « um », hésitations, reformulations). Le doublage restitue globalement un débit adapté, ce qui constitue un point de réussite relatif par rapport aux extraits gaming analysés jusqu'ici. En revanche, le rire de Brian au milieu de sa réplique — qui

	signale qu'il s'amuse lui-même de la situation évoquée et qui crée un moment de complicité avec Albrecht — est entièrement ignoré. La réplique est rendue de façon continue et neutre, sans que ce moment de légèreté ne laisse aucune trace dans le doublage. Les marqueurs d'hésitation oraux (« uh », « um ») disparaissent de la version française, et la double reprise « in in 19 » — dysfluente caractéristique d'un locuteur qui parle de façon spontanée — est lissée en un simple « en 19 ». Ce lissage, déjà observé sur les faux départs de Maya et de Captain Collins, confirme la tendance systémique des systèmes de TAN à normaliser le discours oral vers un français écrit standard, effaçant les marqueurs que Chaume (2012, p. 82) identifie comme constitutifs de l'oralité préfabriquée.
Crédibilité discursive	L'erreur majeure de l'extrait est la prononciation « biochimia » par la voix de synthèse, là où le terme français correct est « biochimie ». Cette erreur est de nature distincte de toutes celles documentées jusqu'ici dans l'analyse : elle n'est pas imputable au module ASR (qui a vraisemblablement transcrit « biochemistry » correctement) ni au module de traduction (qui a produit « biochimie »), mais au module TTS, qui restitue le terme avec une terminaison en « -ia » absente du français. Une hypothèse plausible est que le système de synthèse vocale a appliqué à ce terme un patron de prononciation issu d'une autre langue — l'italien « biochimica », par exemple — ou qu'il a tronqué le mot avant sa syllabe finale. Quelle que soit la cause, le résultat est un terme scientifique rendu incorrectement à l'oral, dans une vidéo de vulgarisation scientifique dont la crédibilité repose précisément sur l'exactitude de la terminologie employée. La formulation « à chaque fois qu'elle le faisait » pour « whenever she was doing it » est fonctionnellement recevable, quoique légèrement lourde pour une construction qui appelait plutôt « quand elle s'y mettait » ou « à l'époque où elle le faisait ». Ce glissement stylistique ne compromet pas la compréhension mais contribue à l'impression d'un discours normalisé plutôt que naturel.
Traduction automatique	Sur les segments correctement transcrits, la traduction est globalement fidèle. La structure argumentative de Brian Greene — la question qu'il aurait voulu poser, le fait d'avoir décroché vingt ans et de devoir rattraper — est restituée sans contresens majeur. L'erreur TTS sur « biochimia » est le seul écart notable sur le plan de la restitution vocale du contenu traduit.
Relation locuteur–spectateur	La configuration à deux intervenants visibles simultanément rend l'uniformisation vocale particulièrement problématique dans ce format. Le spectateur voit en permanence les deux visages — Albrecht et Brian — et entend une seule voix masculine pour les deux. L'attribution d'une voix unique à deux personnes simultanément visibles à l'écran constitue une rupture immédiate de l'incarnation vocale au sens de Bosseaux (2015, p. 66) : il est impossible de construire l'illusion que le corps visible et la voix entendue émanent d'une même présence cohérente, puisqu'un seul des deux visages peut correspondre à la voix à un instant donné. La cohérence de cette uniformisation — la même voix masculine est maintenue pour les deux locuteurs sans instabilité intra-énonciative — ne constitue pas pour autant une réussite : c'est la cohérence d'une erreur systémique, non sa correction.
Bilan de l'extrait	

La défaillance est ici localisée au troisième maillon de la chaîne : « biochimie » est correctement transcrit et traduit, mais mal restitué vocalement. L'extrait établit empiriquement que chacun des trois modules peut produire des erreurs indépendantes, sans défaillance en amont.
Réussite : fidélité sémantique, débit adapté au registre. Lacunes : erreur TTS sur « biochimia », voix masculine unique pour deux locuteurs simultanément visibles, voix originale d'Albrecht audible en fond, rire ignoré, dysfluences normalisées.

FICHE D'ANALYSE — EXTRAIT 2

Vidéo (titre + chaîne)	Straight Talk on Cosmic Origins — World Science Festival
Genre	Vulgarisation scientifique
Horodatage de l'extrait	7:55 – 8:35
Durée de l'extrait	40 s
Numéro d'extrait	2
Transcription	
VO (anglais)	A- so I was there in Paul's office and I was crestfallen to know (rires) he was doing cosmology. B- And did he then, did he explain to you that he was thinking on the verge of ... A- So what, I what I, what happened for me was two things one was I thought he just might know more than I did so, so I was humble enough (rires) to to tune into that point of view and also I was desperate because (rires) I had sort of been using up my chances in different ways and the world had used them up for me. And so I I told him I I swallowed my disappointment I said I'd love to do cosmology with you.
Doublage automatique (FR)	A- Donc j'étais là dans le bureau de Paul et j'étais découragé. Il faisait de la cosmologie. B- Et est-ce qu'il vous a alors expliqué qu'il était sur le point de. A- Donc ce qu'il m'est arrivé, c'est deux choses. La première c'est que je pensais qu'il en savait peut-être plus que moi. (...) Donc j'étais assez humble pour me mettre à l'écoute de ce point de vue et aussi j'étais désespéré parce que j'avais en quelque sorte utilisé mes chances de différente manière et ou le monde les avait utilisées pour moi. (...) Alors je lui ai dit que j'avais ravalé ma déception. J'ai dit : « J'adorerais faire de la cosmologie avec vous. » (débit accéléré)
Observations par critère	
Synchronisation	Deux pauses (...) interrompent le doublage d'Albrecht à des moments où il continue de parler à l'écran. Sa voix originale reste audible en fond sonore durant ces blancs, exposant le décalage entre les deux couches sonores. Ces lacunes sont vraisemblablement causées par les rires et les reformulations d'Albrecht — les dysfluences qui jalonnent son discours dégradent le signal audio au point que le module ASR cesse momentanément de transcrire. L'isochronie est donc rompue à ces deux endroits précis, et l'accélération du débit documentée en fin de réplique confirme le même mécanisme

	compensatoire observé dans les extraits précédents : le système compresse les segments restants pour rattraper le retard accumulé.
Prosodie & expressivité	Dans la VO, la réplique de Brian est délibérément laissée en suspension — « he was thinking on the verge of... » — parce qu’Albrecht l’interrompt avant qu’il n’ait pu la terminer. Cette suspension est marquée prosodiquement : l’intonation monte légèrement et reste ouverte, signalant que l’énoncé est inachevé. Le doublage la restitue avec une intonation déclarative descendante, comme si la phrase était complète. Le spectateur francophone entend une affirmation là où Brian produisait une question-amorce volontairement tronquée. Cette erreur est de nature différente des défaillances de neutralisation prosodique documentées jusqu’ici : le système ne se contente pas d’aplatir une émotion, il substitue une structure intonative à une autre, transformant un énoncé ouvert en énoncé fermé. Les trois rires d’Albrecht sont intégralement ignorés. Leur effacement transforme un récit introspectif et vivant en une succession d’énoncés neutres.
Crédibilité discursive	Le discours d’Albrecht dans cet extrait est dense en marqueurs d’oralité préfabriquée au sens de Chaume (2012, p. 82) : faux départs (« So what, I what I »), répétitions (« so, so », « I I »), hésitations, reformulations. Ces éléments ne sont pas des imperfections, ils sont propres à une personne qui réfléchit à voix haute, qui cherche ses mots. Le doublage les efface systématiquement : « So what, I what I, what happened for me » devient « Donc ce qu’il m’est arrivé », « so, so I was humble enough » devient « j’étais assez humble ». Le résultat est un discours fonctionnellement correct mais prosodiquement vide, qui ne ressemble plus à de la parole réfléchie. L’apparition de « et ou » dans le doublage mérite une analyse spécifique. Dans la VO, Albrecht rit et reprend son fil en disant « and » avant de continuer. Le module ASR, confronté à cette hésitation vocale doublée d’un rire, a vraisemblablement transcrit « and or », produisant un « et ou » sans ancrage syntaxique dans la phrase française. L’erreur est donc due à l’incapacité du module à distinguer un son para-verbal (le rire, la prolongation vocalique) d’un mot lexical. Le résultat est une conjonction double qui rend la phrase syntaxiquement boiteuse pour le spectateur francophone.
Traduction automatique	Sur les segments correctement transcrits, la traduction est fidèle et même stylistiquement soignée sur certains points — « j’avais ravalé ma déception » pour « I swallowed my disappointment » est un rendu idiomatique naturel, qui maintient l’image de la VO sans calque. « Crestfallen » rendu par « découragé » est fonctionnellement correct, quoique légèrement en deçà de la nuance affective du terme anglais — « accablé » ou « atterré » auraient mieux rendu la force émotionnelle du mot original. Ces choix ne constituent pas des erreurs au sens strict, mais ils illustrent la tendance du système à peut-être privilégier l’équivalent le plus courant plutôt que l’équivalent le plus juste stylistiquement.
Relation locuteur–spectateur	L’uniformisation vocale masculine pour les deux intervenants, documentée dans l’extrait 1, se confirme ici sans variation. La configuration à deux visages simultanément visibles à l’écran maintient la même rupture d’incarnation vocale : le spectateur ne peut à aucun moment associer la voix entendue au corps qui lui correspond visuellement. En effaçant les rires, les hésitations, les reprises, le doublage neutralise l’authenticité de l’histoire d’Albrecht et transforme un récit personnel en compte rendu factuel.
Bilan de l’extrait	

Deux fragilités situées hors du champ des erreurs terminologiques habituelles. Une intonation déclarative transforme un énoncé volontairement suspendu en affirmation close. Et le prolongement vocalique d'une hésitation est transcrit comme un mot lexical, produisant le « et ou » incohérent. C'est la dimension prosodique et para-verbale qui est ici en cause, à toutes les étapes de la chaîne.

Réussite : rendu idiomatique de « swallowed my disappointment ». Lacunes : intonation fautive sur une réplique suspendue, « et ou » issu d'une hésitation, trois rires ignorés, dysfluences normalisées.

FICHE D'ANALYSE — EXTRAIT 3	
Vidéo (titre + chaîne)	Straight Talk on Cosmic Origins — World Science Festival
Genre	Vulgarisation scientifique
Horodatage de l'extrait	10:55 – 11:40
Durée de l'extrait	45 s
Numéro d'extrait	3
Transcription	
VO (anglais)	B- And that motivated Allen to write that paper that showed up in your fortuitous Phys Rev D copy. Can (stutter), I don't know if we want to go through all of them, but I think a big one especially to motivate inflation is is trying to understand the horizon problem, you know? Why one part of the sky is so similar to the other. [Yeah, please go ahead.] A- By the way, I can also slip in that Allen um refers to a talk at Cornell that influenced him to think about these problems. B- Bob Dickey talk, right? A- Yes, yes, um and and I was, I I was at Cornell at the time, I I had was sort of curious to go to that lecture, but something had happened um to get in the way.
Doublage automatique (FR)	B- Et cela motivait Alan à écrire cet article qui était apparu dans l'année pour qu'il soit (xxx) des copies, je veux dire, je ne sais pas si nous voulons tous (sans liaison) les passer en revue, mais je pense qu'un des plus importants, en particulier pour motiver l'inflation, est d'essayer de comprendre le problème de l'horizon. Vous savez pourquoi une partie du ciel est si semblable à l'autre ? A- Au fait, je peux aussi glisser qu'Alan fait référence à une conférence à Cornell qui l'a influencé à réfléchir à ces problèmes. B- Bob Dickey parle bien. A- Oui, oui, et j'étais à Cornell à l'époque (...) j'étais un peu curieux d'aller à une conférence, mais quelque chose s'était passé qui m'en a empêché.
Observations par critère	
Synchronisation	La voix originale d'Albrecht reste audible en fond sonore sur l'ensemble de l'extrait, interférant régulièrement avec le doublage — phénomène déjà documenté dans les extraits précédents de cette vidéo. La pause (...) dans la réplique finale d'Albrecht signale un nouveau blanc de transcription, durant lequel il continue de parler à l'écran sans voix correspondante dans le doublage. Le chevauchement entre Brian et Albrecht, « Yeah, please go ahead », est entièrement ignoré, le module ASR ayant vraisemblablement écarté cette intervention brève au profit du discours de Brian, sans que cette sélection soit marquée d'aucune façon dans la version doublée.

Prosodie & expressivité	Le bégaiement de Brian (« Can (stutter) ») est rendu par « je veux dire » — marqueur discursif courant qui ne correspond à aucun équivalent du discours d'origine. Le système n'a pas transcrit le bégaiement comme tel, et a produit un remplacement générique qui lisse l'hésitation en une transition neutre. Ce traitement confirme la tendance à normaliser les dysfluences orales documentée depuis l'extrait 2. L'intonation de la question de Brian — « Why one part of the sky is so similar to the other? » — est restituée correctement comme interrogative dans le doublage, ce qui constitue un point de réussite partiel par rapport à l'extrait 2, où la réplique suspendue de Brian avait été rendue avec une intonation déclarative.
Crédibilité discursive	Trois écarts distincts. « Phys Rev D », abréviation courante de Physical Review D, n'est pas reconnue et produit un blanc : dans une vidéo portant sur la physique théorique, l'effacement d'une référence bibliographique centrale prive le spectateur d'une information du discours. « Tous » est prononcé sans liaison, erreur de synthèse vocale de même nature que « biochimia ». Enfin, « a talk at Cornell, right ? » devient « Bob Dickey parle bien » : « talk » est traité comme un verbe et « right ? » rendu par « bien ». Le résultat est syntaxiquement recevable en français — ce qui le rapproche de « je ne suis pas malin » — mais sémantiquement sans rapport avec la VO, où Brian demande confirmation. La réplique perd sa fonction de relance du récit d'Albrecht.
Traduction automatique	Cet extrait concentre des erreurs à deux niveaux de la chaîne de traitement. Au niveau de l'ASR : non-reconnaissance de « Phys Rev D » et mauvaise segmentation de « Bob Dickey talk, right ? ». Au niveau du TTS : absence de liaison sur « tous les ». Sur les segments non affectés par ces erreurs, la traduction est correcte : « le problème de l'horizon », « une partie du ciel est si semblable à l'autre », « j'avais ravalé ma déception » — rendu ici repris de façon cohérente — montrent que le système produit des traductions fluides lorsque le discours est simple et le signal audio clair.
Relation locuteur–spectateur	La réplique « Bob Dickey parle bien » illustre une rupture spécifique de la continuité communicative. Dans la VO, Brian demande une confirmation — « Bob Dickey talk, right ? » — qui sert à la fois à relancer le récit d'Albrecht et à montrer qu'il connaît l'histoire et y participe activement. C'est une intervention qui signale à l'audience que Brian n'est pas un animateur passif mais il est informé. La transformation de cette réplique en commentaire neutre sur la capacité de Bob Dickey à bien parler efface entièrement la dynamique d'échange entre les deux intervenants. Le spectateur francophone ne comprend pas pourquoi Brian dit cela, ni en quoi c'est pertinent à ce moment de la conversation.
Bilan de l'extrait	
L'erreur « Bob Dickey parle bien » appartient à la même catégorie que « je ne suis pas malin » : un énoncé grammaticalement correct et pragmatiquement recevable hors contexte, mais sémantiquement sans rapport avec l'original. Ces erreurs discrètes sont plus coûteuses que les erreurs aberrantes, car elles ne déclenchent aucune alerte et induisent une compréhension fautive. Réussite : intonation interrogative correcte sur la question de Brian. Lacunes : « Phys Rev D » non reconnu, segmentation erronée sur « Bob Dicke talk », liaison manquante sur « tous les », bégaiement normalisé, voix originale audible en fond.	

FICHE D'ANALYSE — EXTRAIT 4	
Vidéo (titre + chaîne)	Straight Talk on Cosmic Origins — World Science Festival

Genre	Vulgarisation scientifique
Horodatage de l'extrait	15:50 – 16:24
Durée de l'extrait	34 s
Numéro d'extrait	4
Transcription	
VO (anglais)	A- You know I could say, well you know I just got back from a trip, and there was this guy sitting next to me on the plane. And of all of billions of people in the world, it was that guy! (rires) How could that be? (rires) And of course there's gotta be some guy, you know (débit accéléré, voix plus aiguë). B- Right. A- So, so, so it's the capture that flavour of like, should I, should I worry about it, or not? What, what, what is an, what is an important um, coincidence to puzzle about and what is not?
Doublage automatique (FR)	A- Vous savez, je reviens juste d'un voyage. Et il y avait ce type assis à côté de moi dans l'avion, et parmi tous les milliards de personnes dans le monde, c'était ce type. (xxx) Comment est-ce possible ? Et bien sûr, il doit y avoir un type, vous savez ? Donc. Donc. Donc, cela capture cette saveur du genre dois-je m'en inquiéter ou pas ? Qu'est-ce (.) que c'est qu'est-ce qui est important euh une coïncidence à propos de laquelle il faut réfléchir et qu'est-ce qui ne l'est pas ?
Observations par critère	
Synchronisation	Deux moments de rupture synchronique sont documentables sur cet extrait. Le premier correspond au passage à voix plus aiguë d'Albrecht — le changement de registre vocal est suffisamment marqué pour que la voix originale remonte au premier plan de la bande sonore, coexistant avec la voix synthétique plutôt qu'en se laissant masquer. Le second est la pause (xxx) dans le doublage, pendant laquelle la voix d'Albrecht reste audible — sûrement liée à ses rires et changements d'intonation, qui interrompent le flux vocal. La réplique brève de Brian (« Right. ») est entièrement absente du doublage, le système a vraisemblablement traité la séquence comme un monologue d'Albrecht.
Prosodie & expressivité	L'extrait repose sur une anecdote qui débouche sur la nature des coïncidences. Le changement d'intonation (« it was that guy! »), les rires qui ponctuent l'anecdote, l'accélération du débit sont impératifs dans cette anecdote. Le doublage les neutralise intégralement : les rires disparaissent, l'accélération est rendue de façon mécanique, et la hausse de la voix, qui dans la VO produit un effet comique, est précisément le moment où la voix originale ressort, trahissant le doublage. La triple répétition « So, so, so », léger bégaiement d'Albrecht qui dans la VO marque la conclusion de son anecdote, est rendue par trois « Donc. » successifs, chacun traité comme une phrase indépendante avec une intonation conclusive. Le résultat est une séquence hachée, sans cohérence.
Crédibilité discursive	Deux phénomènes. Le premier est la restitution du « um » d'Albrecht par un « euh », cas presque inverse de ceux documentés jusqu'ici : là où le système efface habituellement les marqueurs para-verbaux, il en produit un, mal placé, entre « important » et « une coïncidence ». Le résultat est plus artificiel que ne l'aurait été son effacement, parce que ce « euh » n'est ancré dans aucun corps : il sonne comme un bruit inséré, non comme une pensée en formation. Le second est la restitution partielle des faux départs de la question finale, où « qu'est-ce » apparaît deux fois séparés d'une pause brève : la fragmentation évoque un défaut de lecture plutôt qu'une hésitation humaine. À l'inverse, la triple anaphore « So, so, so » est maintenue, et « like » est rendu par « du genre », équivalent idiomatique juste.

Traduction automatique	Sur les segments correctement transcrits, la traduction est fidèle. « Of all of billions of people in the world, it was that guy » rendu par « parmi tous les milliards de personnes dans le monde, c'était ce type » est une transposition directe et fonctionnelle. « Capture that flavour » rendu par « capture cette saveur » maintient la métaphore de l'original sans calque maladroit. « Coincidence to puzzle about » rendu par « coïncidence à propos de laquelle il faut réfléchir » est sémantiquement correct, quoique légèrement lourd — « coïncidence sur laquelle se pencher » aurait produit un résultat plus naturel à l'oral. « Should I, should I worry » est rendu par « dois-je m'en inquiéter » sans la répétition — perte mineure mais cohérente avec la tendance à la normalisation documentée depuis l'extrait 1.
Relation locuteur–spectateur	L'anecdote du voisin d'avion ancre une question qui peut-être abstraite dans une expérience quotidienne pour la rendre accessible. En ignorant, les rires et intonations changeantes d'Albrecht, le doublage transforme un récit vivant en exposé factuel.
Bilan de l'extrait	
Deux phénomènes inédits. Le filtre d'atténuation ne résiste pas à un pic de hauteur vocale et laisse percer la voix d'Albrecht. Et la restitution d'un marqueur d'hésitation (« um » rendu « euh ») produit un résultat plus artificiel que son effacement : la voix de synthèse reproduit la forme sonore de l'hésitation sans pouvoir l'incarner, ce qui suffit à trahir l'absence de corps. Réussite : traduction fidèle, « like » rendu « du genre ». Lacunes : perçage de la VO sur le pic d'intonation, deux rires ignorés, réplique de Brian effacée, « euh » synthétique parasite, faux départs rendus de façon incohérente.	

Big Think - Brian Cox on how black holes could unlock the mysteries of our universe

FICHE D'ANALYSE — EXTRAIT 5	
Vidéo (titre + chaîne)	Brian Cox on how black holes could unlock the mysteries of our universe — Big Think
Genre	Vulgarisation scientifique
Horodatage de l'extrait	1:06 – 1:20
Durée de l'extrait	14 s
Numéro d'extrait	5
Transcription	
VO (anglais)	I am Brian Cox. I am a professor of particle physics at the University of Manchester and the Royal Society Professor for the Public Engagement in Science, and amongst other things, I am the author of the book « Black Holes: the key to understanding the universe. »
Doublage automatique (FR)	Je suis Brian Cox. Je suis professeur de physique des particules à l'université de Manchester et professeur de la Royal Society pour l'engagement du public dans la science, et je suis, entre autres, l'auteur du livre « Black holes: the key to understanding the universe. » (Accent français très prononcé)
Observations par critère	

Synchronisation	L'isochronie est globalement respectée sur l'ensemble de la séquence. Cox est visible à l'écran et son débit posé facilite l'alignement entre le flux original et le doublage. Aucun blanc notable, aucune accélération compensatoire.
Prosodie & expressivité	Le registre de cet extrait est celui d'une présentation formelle et maîtrisée — Cox récite ses titres et sa présentation de façon posée, sans variation émotionnelle marquée. Le doublage restitue un débit adapté à ce registre. L'absence de marqueurs expressifs complexes dans la VO limite les défaillances prosodiques potentielles : cet extrait constitue, parmi tous ceux analysés dans l'ensemble du corpus, le contexte le plus favorable au système.
Crédibilité discursive	La première observation sur ce critère est une réussite notable et peu fréquente dans le corpus : le système a correctement identifié les éléments qui ne devaient pas être traduits. « Royal Society », nom propre d'une institution britannique, est maintenu en anglais. Le titre du livre est conservé dans sa langue originale. « University of Manchester » est maintenu tel quel. Ces choix témoignent d'une forme de compétence implicite du système à reconnaître certaines entités nommées comme intraduisibles — compétence que les extraits gaming avaient systématiquement mise en défaut sur les noms propres de jeux et de champions. La seconde observation concerne la prononciation du titre du livre. Bien que les mots anglais soient correctement conservés à l'écrit, la voix de synthèse les restitue avec un accent français très prononcé. Cette erreur est sûrement due au module TTS, qui applique les règles phonologiques du français à une séquence lexicale anglaise. Il n'y aurait donc pas de mécanisme de changement de langue au sein d'un même énoncé. Le système reconnaît que le texte doit être conservé en anglais, mais ne peut pas restituer une phonologie anglaise depuis une voix de synthèse paramétrée en français.
Traduction automatique	La traduction des segments effectivement rendus en français est précise et stylistiquement correcte. « Professor for the Public Engagement in Science » rendu par « professeur pour l'engagement du public dans la science » est un calque fonctionnel et non erroné. « Amongst other things » rendu par « entre autres » est naturel. Aucune erreur sémantique n'est détectable.
Relation locuteur–spectateur	En se présentant, Brian Cox établit sa légitimité académique et présente son ouvrage. Ces deux fonctions sont préservées dans le doublage, le spectateur francophone comprend qui est Cox et de quoi il sera question. La prononciation approximative du titre du livre en accent français est la seule entorse à la crédibilité de cette présentation.
Bilan de l'extrait	
Extrait ciblé sur le résultat le plus réussi du corpus. Il documente ce que le système produit en conditions optimales : locuteur unique, débit posé, registre formel, ni jargon non reconnu ni chevauchements. Réussite : fidélité sémantique totale, entités intraduisibles conservées (Royal Society, titre de l'ouvrage, University of Manchester), isochronie respectée, registre maintenu. Lacune : phonologie française appliquée au titre anglais.	

FICHE D'ANALYSE — EXTRAIT 6	
Vidéo (titre + chaîne)	Brian Cox on how black holes could unlock the mysteries of our universe — Big Think
Genre	Vulgarisation scientifique

Horodatage de l'extrait	1:46 – 2:23
Durée de l'extrait	37 s
Numéro d'extrait	6
Transcription	
VO (anglais)	One was a clergyman called Mitchell, and the other was the great French mathematician, Laplace. And they were both thinking in terms of an idea called "escape velocity". The escape velocity is the speed you have to travel to completely escape the gravitational pull of something, a planet or a star. So for the earth, for example, the escape velocity from the surface of the earth is around 8 miles a second. If you go bigger, you make a bigger, more massive thing, let's go to a star, for example, like the sun, it's somewhere in the region of 400 miles a second.
Doublage automatique (FR)	L'un était un ecclésiastique appelé Mitchell et l'autre était le grand mathématicien français liasse (débit accéléré). Ils réfléchissaient tous les deux en termes de vitesse de libération. La vitesse de libération est la vitesse à laquelle il faut se déplacer pour échapper complètement au pôle gravitationnel d'une planète ou d'une étoile. Ainsi, pour la terre, par exemple, la vitesse de libération depuis la surface de la terre est d'environ 8 mètres par secondes. Si vous agrandissez vous créez une chose plus grande et plus massive. Prenons l'exemple d'une étoile, comme le soleil, qui est d'environ 640 kilomètres par secondes.
Observations par critère	
Synchronisation	L'isochronie est globalement respectée, à l'exception d'une légère accélération du débit lors de la mention de Laplace.
Prosodie & expressivité	La scission de la dernière phrase crée un léger problème de compréhension. Dans la VO, « If you go bigger, you make a bigger, more massive thing, let's go to a star, for example, like the sun, it's somewhere in the region of 400 miles a second » est une phrase continue. Le doublage coupe cette phrase en deux énoncés indépendants : « Si vous agrandissez vous créez une chose plus grande et plus massive. » puis « Prenons l'exemple d'une étoile, comme le soleil, qui est d'environ 640 kilomètres par seconde. » Le premier énoncé est rendu avec une intonation déclarative descendante, comme s'il était complet en lui-même. Le spectateur francophone reçoit donc d'abord une affirmation générale — qui sonne comme une règle posée pour elle-même —, puis un exemple qui semble débiter un nouveau développement plutôt que d'illustrer ce qui précède. La continuité argumentative de Cox est rompue.
Crédibilité discursive	Deux écarts affectent la crédibilité de l'exposé. « Laplace » est correctement transcrit mais restitué « Liasse » par la synthèse vocale, erreur de même nature que « biochimia ». Et la vitesse de libération terrestre, que Cox énonce en miles par seconde, est restituée en mètres par seconde : la donnée s'en trouve abaissée de plusieurs ordres de grandeur. Dans un contenu dont la fonction première est la transmission de connaissances exactes, une donnée chiffrée fautive et un nom de savant méconnaissable atteignent directement l'autorité du locuteur. La séquence convertit pourtant correctement les « 400 miles a second » suivants, ce qui montre que le système dispose du mécanisme et l'applique de façon inconstante.
Traduction automatique	Deux réussites de la traduction automatique méritent d'être relevées. « Escape velocity » est rendu par « vitesse de libération », qui est le terme technique français correct. « Clergyman » rendu par « ecclésiastique » est un choix lexical précis et adapté au registre de la vidéo. Les erreurs documentées ci-dessus — « miles » → « mètres », « pull » → « pôle », « Laplace » → « Liasse » — illustrent que les réussites et les échecs existent dans le même extrait, sans logique apparente de cohérence. Le

	système produit « vitesse de libération » avec exactitude et « pôle gravitationnel » avec erreur dans la même phrase, sans que la proximité des deux termes lui permette de corriger le second à partir du premier.
Relation locuteur–spectateur	Cox construit son explication en s’adressant directement à son audience, « let’s go to a star, for example, like the sun », et en progressant par paliers. La scission de la phrase finale coupe précisément le moment où Cox invite son audience à le suivre dans ce passage par paliers. « Si vous agrandissez vous créez une chose plus grande et plus massive », rendu comme une affirmation autonome, perd la dimension d’invitation et de progression pédagogique que la VO portait.
Bilan de l’extrait	
Deux erreurs situées à des maillons différents. Une substitution ASR rend « miles » par « mètres » et produit une donnée fautive de plusieurs ordres de grandeur, dans un contenu dont la fonction première est la transmission de connaissances exactes. Une erreur TTS déforme « Laplace » en « Liasse », de même nature que « biochimia ». La conversion correcte des « 400 miles » en kilomètres dans la même séquence montre que le système dispose du mécanisme et l’applique de façon inconstante. Réussite : « escape velocity » rendu « vitesse de libération », « clergyman » rendu « ecclésiastique ». Lacunes : « Laplace » → « Liasse », « miles » → « mètres », « gravitational pull » → « pôle gravitationnel », scission de la phrase conditionnelle.	

FICHE D’ANALYSE — EXTRAIT 7	
Vidéo (titre + chaîne)	Brian Cox on how black holes could unlock the mysteries of our universe — Big Think
Genre	Vulgarisation scientifique
Horodatage de l’extrait	11:59 – 12:18
Durée de l’extrait	19 s
Numéro d’extrait	7
Transcription	
VO (anglais)	Einstein said that if you look at nature really carefully and keep pulling at the intellectual threads and keep going and just keep delving down into what nature seems to be trying to tell us, then if you’re lucky and persistent, you can catch a glimpse of something deeply hidden.
Doublage automatique (FR)	Einstein a dit que si vous observez la nature très attentivement et que vous continuez à tirer sur les fils (prononcé fis) intellectuels, à continuer et à creuser ce que la nature semble essayer de nous dire. (..) Si vous avez de la chance et que vous persévérez, vous pourrez apercevoir quelque chose de profondément caché.
Observations par critère	
Synchronisation	Cox n’est pas visible à l’écran sur cet extrait, il s’exprime en voix off sur des images. La synchronisation labiale n’est donc pas un enjeu ici, ce qui place le doublage dans des conditions a priori favorables. Cependant, la pause (..) introduite au milieu de la phrase montre que le système a découpé l’énoncé en deux unités distinctes à un endroit où une segmentation n’était pas nécessaire.

Prosodie & expressivité	Cox parle délibérément lentement sur cette citation d'Einstein pour marquer son importance auprès de l'audience. Le doublage restitue un débit assez rapide et fragmente la citation en deux affirmations ordinaires.
Crédibilité discursive	Une erreur importante est la prononciation de « fils », pluriel de « fil », comme /fis/, c'est-à-dire avec la prononciation du mot « fils » désignant un enfant masculin. Le module TTS a sûrement appliqué la prononciation statistiquement dominante du mot « fils », plus fréquent dans les corpus écrits — sans tenir compte du contexte syntaxique, qui indiquait clairement qu'il s'agissait du pluriel de « fil » (« tirer sur les fils intellectuels »). Cette confusion produit une image absurde là où la métaphore de Cox était claire dans la VO.
Traduction automatique	La traduction est globalement fidèle sur le plan sémantique. « Keep pulling at the intellectual threads » est rendu par « tirer sur les fils intellectuels » — métaphore correctement transposée, malgré l'erreur de prononciation documentée ci-dessus. « Keep delving down » devient « creuser », choix idiomatique naturel. « Catch a glimpse » est rendu par « apercevoir », équivalent fonctionnel correct. « Deeply hidden » devient « profondément caché », calque direct mais acceptable. Le système produit ici une traduction de qualité sur le fond — les choix lexicaux sont justes et la métaphore d'Einstein est préservée. C'est la restitution vocale qui trahit ce que la traduction avait correctement produit.
Relation locuteur-spectateur	L'absence de Cox à l'écran modifie le rapport au spectateur : sans corps visible, la voix devient l'unique vecteur de présence. Dans la VO, avec son intonation, Cox invite son audience à s'arrêter sur cette citation, à en ressentir l'importance. Le doublage neutralise entièrement cette invitation.
Bilan de l'extrait	
Le module TTS ne distingue pas deux mots graphiquement identiques dont la prononciation dépend du contexte syntaxique. L'extrait montre en outre que l'absence du locuteur à l'écran, condition a priori favorable puisqu'elle supprime la contrainte labiale, ne suffit pas : c'est ici la voix seule qui doit porter le sens. Comme le souligne Bosseaux (2015, p. 66), la voix ne transmet pas seulement un contenu, elle construit une présence — et sans corps derrière elle, elle ne construit rien. Réussite : métaphore des fils correctement transposée, « catch a glimpse » rendu « apercevoir ». Lacunes : désambiguïsation homographique fautive sur « fils », segmentation de la phrase conditionnelle, neutralisation prosodique d'une citation mise en valeur par Cox.	

9.3 Synthèse du doublage automatique face aux vidéos de vulgarisation scientifique

L'analyse des sept extraits du sous-corpus de vulgarisation scientifique révèle un profil de défaillance distinct de celui documenté dans le gaming, non pas parce que le système se comporte mieux, mais parce qu'il échoue à des endroits différents de la chaîne de traitement et

pour des raisons qui tiennent à la nature du discours scientifique plutôt qu'à sa complexité sonore.

La défaillance la plus caractéristique de ce sous-corpus se situe au niveau du module TTS. Contrairement aux erreurs ASR qui dominaient le corpus gaming, les extraits de vulgarisation scientifique révèlent un module de synthèse vocale qui produit des erreurs sur des termes pourtant correctement transcrits et traduits : « biochimie » restitué « biochimia », « Laplace » restitué « Liasse », « fils » au sens de fil prononcé comme le mot de parenté. Ces trois erreurs affectent le dernier maillon de la chaîne, celui qui produit le signal sonore entendu par le spectateur, sans que les étapes précédentes soient en cause. Elles révèlent que le module TTS traite certains mots comme des séquences phonétiques ordinaires, sans ancrer leur prononciation dans une norme conventionnelle : « Laplace » est segmenté de façon erronée, « fils » reçoit la prononciation statistiquement dominante de ses deux acceptions homographes, et le titre anglais de l'ouvrage de Cox est prononcé avec une phonologie entièrement française. Ce dernier cas documente une limite architecturale précise : le système ne dispose d'aucun mécanisme de commutation phonologique entre le français et l'anglais à l'intérieur d'un même énoncé.

La segmentation erronée des phrases longues constitue la seconde constante du sous-corpus, et elle se manifeste de façon régulière. Sur trois extraits distincts (2, 6 et 7), une même configuration produit le même résultat : une phrase conditionnelle ou rhétorique construite sur une progression (si... alors, ou si... c'est parce que) est découpée en deux énoncés indépendants par le système, qui attribue une intonation déclarative descendante au premier segment. Le spectateur reçoit une affirmation close là où l'orateur produisait une hypothèse ouverte appelant une résolution. Ce phénomène n'est imputable ni à un défaut de transcription ni à un défaut de traduction, le contenu lexical étant fidèle. C'est le découpage prosodique appliqué par le module TTS aux segments traduits qui transforme la structure rhétorique en succession d'énoncés formellement autonomes. Comme le souligne Bosseaux, la voix ne transmet pas seulement un contenu : elle construit une présence et une intention (2018). En segmentant de façon mécanique, le doublage automatique rend le contenu sans en porter le sens.

Les marqueurs para-verbaux du discours spontané — hésitations, rires, prolongations vocaliques — constituent un troisième point de fragilité, spécifique à la vidéo du World Science Festival. L'interview entre Brian Greene et Andreas Albrecht est dense en ces éléments, qui

dans la VO remplissent des fonctions précises : les rires ponctuent les moments de connivence intellectuelle, les hésitations signalent qu'un locuteur formule une pensée en temps réel, les reprises marquent une autocorrection. Le module ASR les traite comme des perturbations du signal plutôt que comme des actes communicationnels : les rires provoquent des blancs de non-transcription pendant lesquels Albrecht reste visible à l'écran, et une prolongation vocalique sur « and » est transcrite comme « and or », produisant le « et ou » syntaxiquement incohérent de l'extrait 2. La voix originale d'Albrecht, maintenue à un volume réduit en fond sonore tout au long de la vidéo, signale par ailleurs que le doublage n'efface pas le locuteur source — il le recouvre imparfaitement, laissant le spectateur percevoir simultanément deux couches sonores qui ne se correspondent pas.

Le format en écran scindé du World Science Festival impose au doublage une contrainte d'incarnation vocale qu'il ne peut pas satisfaire. Les deux intervenants sont simultanément visibles tout au long de la vidéo, ce qui signifie que le spectateur voit en permanence deux corps distincts auxquels une seule voix masculine est attribuée. L'incarnation vocale suppose que la voix cible crée l'illusion que le corps visible et la voix entendue émanent d'une même présence (Bosseaux 2015, p. 66). Ici, cette illusion est structurellement impossible à construire : quelle que soit la qualité de la voix synthétique, elle ne peut correspondre qu'à l'un des deux visages à la fois, et le spectateur le perçoit immédiatement. Comme le rappelait le chapitre 2 à propos du cas Jean-Pierre Michaël, l'uniformisation vocale n'est pas propre au doublage automatique. La différence tient à son statut : choix économique et de casting dans le doublage professionnel, où l'appariement entre une voix et un corps demeure stable à l'intérieur d'une même production, elle résulte ici de l'absence de tout module d'identification des locuteurs. Elle n'est donc pas négociée, mais subie.

L'extrait 5, seul extrait ciblé du sous-corpus, documente le résultat le plus satisfaisant de l'ensemble du corpus analysé. Sur la présentation de Brian Cox, le doublage restitue fidèlement le contenu sémantique, maintient correctement les entités intraduisibles (Royal Society, titre de l'ouvrage) et respecte l'isochronie. Cette réussite relative s'explique par la conjonction de plusieurs facteurs favorables : un locuteur unique, un débit posé, une syntaxe simple, un registre prévisible. Elle montre que le système produit un transfert linguistique acceptable lorsque le discours s'approche des conditions pour lesquelles il a été conçu, à savoir un texte normalisé, continu et monologique. L'erreur TTS sur la phonologie du titre anglais, qui survient sur le seul

élément étranger de la séquence, illustre que même dans ces conditions optimales, la limite du module de synthèse vocale persiste dès qu'il sort du français standard.

Sur l'ensemble du sous-corpus, le doublage automatique parvient à transmettre la structure informative générale des exposés : les arguments de Brian Greene et les concepts physiques expliqués par Cox restent globalement décodables. C'est une différence notable avec les contenus gaming les plus complexes, où le sens pouvait être entièrement perdu. Mais cette transmission informationnelle partielle ne suffit pas à constituer un doublage réussi au sens de Marzà et Chaume, pour qui ce qui importe est l'impression produite et la crédibilité de la parole saisie comme un tout (2009, p. 35). Un exposé scientifique ne se réduit pas à son contenu propositionnel : il repose sur une voix qui construit une démonstration en s'adressant à un public. Darwish et Orero ont montré, à partir d'un reportage de la BBC, que le voice-over peut transmettre une information qui ne correspond pas à l'original tout en conservant l'apparence de fidélité que lui donne la présence audible de la voix source (2014, cité dans Matamala 2018, p. 72). Le mécanisme observé ici est comparable dans ses effets : le spectateur entend la voix de Cox sous le doublage, ce qui garantit à ses yeux l'authenticité du propos, alors même que la donnée chiffrée qui lui parvient est fausse. Il en diffère cependant sur un point décisif. Chez Darwish et Orero, la distorsion procède d'un choix éditorial ; ici, elle procède d'une défaillance technique que personne n'a décidée et que personne ne contrôle.

PARTIE IV – RÉSULTATS ET DISCUSSION

10. Comparaison intergenres

Les deux sous-corpus ayant été analysés séparément, il devient possible de les confronter. Ce chapitre identifie ce que les deux genres ont en commun, ce qui les distingue, et montre que les cinq critères de la grille n'ont pas le même poids selon les conventions du genre considéré.

10.1 Points de convergence entre gaming et vulgarisation scientifique

Malgré des configurations communicationnelles très différentes, les deux sous-corpus font apparaître des défaillances communes. Plutôt que de les énumérer à nouveau, on cherchera ici à déterminer ce qui les rend transversales, c'est-à-dire ce qu'elles partagent au-delà de leurs manifestations particulières.

Les quatre constantes relevées dans les synthèses précédentes — neutralisation prosodique, effacement des marqueurs para-verbaux, uniformisation vocale, persistance audible de la voix originale — procèdent d'une même propriété du dispositif : la chaîne de traitement ne comporte aucune représentation du sujet parlant. Elle traite un flux audio, en extrait une suite de mots, la transpose dans une autre langue et la restitue au moyen d'une voix disponible. À aucun moment elle ne modélise qui parle, à qui, ni dans quelle intention. La neutralisation prosodique en découle parce que l'intonation n'est pas reconstruite à partir d'une intention mais générée à partir d'un texte ; l'effacement des marqueurs para-verbaux en découle parce que rires et hésitations ne sont pas reconnus comme des actes mais détectés comme des perturbations du signal ; l'uniformisation vocale en découle parce qu'aucun module n'associe une voix à un locuteur ; et la persistance de la voix originale en découle parce que l'atténuation est appliquée de façon uniforme au lieu d'être modulée selon la structure des tours de parole.

Cette unité de cause a une conséquence importante pour l'appréciation des résultats. Elle impose de distinguer deux ordres de défaillance. Certaines sont contingentes et susceptibles d'être corrigées par un enrichissement des données ou des ressources : les erreurs de reconnaissance portant sur le jargon vidéoludique, la prononciation fautive des noms propres français, l'absence de commutation phonologique entre le français et l'anglais. D'autres sont architecturales : l'absence d'identification des locuteurs, l'absence d'incarnation vocale, l'absence de toute étape de révision humaine. Les premières relèvent d'une performance insuffisante et se résorberont vraisemblablement ; les secondes relèvent de la conception du

dispositif et ne se résorberont pas sans que celle-ci soit modifiée. Cette distinction structure l'ensemble de la discussion qui suit.

Un dernier point de convergence mérite d'être noté, qui ne relève pas des défaillances mais des réussites. Dans les deux sous-corpus, l'isochronie est globalement respectée et la fidélité sémantique est satisfaisante sur les segments explicatifs simples. Le système accomplit correctement ce qu'il est conçu pour accomplir, à savoir un transfert informatif ajusté à la durée du segment source. Les défaillances relevées ne portent donc pas sur l'exécution de sa tâche, mais sur l'écart entre cette tâche et ce qu'exige le doublage.

Ces observations valent pour les deux genres. Elles n'épuisent cependant pas les résultats, car les deux sous-corpus se distinguent nettement par la localisation des erreurs dans la chaîne, par le degré de dégradation atteint et par la nature des conventions mises en défaut.

10.2 Points de divergence

Si les convergences révèlent des limites structurelles du dispositif, les divergences entre genres éclairent la façon dont ces limites interagissent avec la nature spécifique de chaque type de discours.

La localisation des erreurs dans la chaîne de traitement diffère significativement entre les deux sous-corpus. Dans le gaming, les défaillances les plus fréquentes et les plus graves se situent au niveau de l'ASR : le jargon hyperspécialisé de *League of Legends* ou d'*Elden Ring* est quasi absent des corpus d'entraînement des systèmes de reconnaissance vocale, et les chevauchements constants dégradent le signal au point de rendre la transcription erronée ou incomplète. Dans la vulgarisation scientifique, le vocabulaire technique de base — termes cosmologiques, mécaniques célestes — est souvent correctement transcrit et traduit, car il est représenté dans les corpus d'entraînement. Les erreurs récurrentes se situent en revanche au niveau du module TTS : prononciation incorrecte de noms propres français (Laplace → Liasse), désambiguïsation homographique erronée (« fils », au sens de fil, prononcé [fis]), phonologie française appliquée à des mots anglais conservés tels quels. La chaîne de traitement est donc vulnérable à des maillons distincts selon le genre — ce qui implique que les stratégies d'amélioration du système ne seraient pas les mêmes d'un corpus à l'autre.

Le degré de dégradation communicationnelle n'est pas non plus équivalent. Dans le sous-corpus gaming, la vidéo de Doael produit un résultat dont aucune des fonctions communicatives n'est préservée : le jargon est méconnaissable, les locuteurs sont indistincts, le fil de l'échange est perdu. Dans la vulgarisation scientifique, même les extraits les plus défaillants restent partiellement décodables : le spectateur francophone comprend qu'Andreas Albrecht parle de sa carrière et que Brian Cox explique la vitesse de libération, même si la précision terminologique et la texture rhétorique sont perdues. Cette différence de degré tient à la densité de l'expressivité orale et à la complexité de la configuration sonore, les deux variables identifiées dans la synthèse gaming. La vulgarisation scientifique produit un discours moins dense en chevauchements et en jargon irréductible, ce qui préserve un seuil minimal de lisibilité.

Les conventions discursives propres à chaque genre exposent enfin des vulnérabilités distinctes. Le gaming sur YouTube est structuré autour de pratiques communicationnelles codifiées — remerciements nominatifs aux abonnés, interpellations directes de l'audience, vocabulaire communautaire (« cringe », « sub », « gg ») — que le système ne reconnaît pas comme telles et neutralise. La vulgarisation scientifique repose sur des conventions d'autorité épistémique — attribution des idées à leurs auteurs, précision des références, cohérence argumentative — que le système compromet par des substitutions terminologiques (« Bob Dicke talk » rendu « Bob Dickey parle bien ») ou des erreurs factuelles graves. L'exemple le plus net concerne la vitesse de libération terrestre : Cox énonce « eight miles a second », soit environ 12,9 kilomètres par seconde — valeur qu'il arrondit lui-même vers le haut, la vitesse réelle avoisinant 11,2 kilomètres par seconde, c'est-à-dire quelque sept miles par seconde. Le doublage restitue « 8 mètres par seconde », substituant l'unité et abaissant la donnée d'un facteur supérieur à mille cinq cents. La même séquence convertit pourtant correctement les « 400 miles a second » de la phrase suivante en « 640 kilomètres par seconde », ce qui montre que le système dispose bien d'un mécanisme de conversion et que la défaillance tient à son application inconstante. Dans les deux genres, l'atteinte porte sur ce qui fonde le pacte implicite entre le créateur et son audience : la confiance communautaire d'un côté, la crédibilité intellectuelle de l'autre.

10.3 Hiérarchisation des critères selon les genres

L'analyse intergenres ne montre pas seulement que les défaillances du doublage automatique varient selon le type de contenu : elle révèle que les cinq critères de la grille analytique n'ont pas le même poids selon le genre, parce que les actes communicationnels qu'ils mesurent n'occupent pas la même place dans chaque type de discours. Cette variation n'est pas propre au doublage automatique : Chaume observe que les différents genres audiovisuels appellent des degrés de synchronisation différents, le cinéma exigeant une perfection que la série télévisée ne requiert pas au même degré (2012, p. 76, cité dans Bosseaux 2018, p. 52).

Dans le gaming, la prosodie et l'expressivité constituent le critère le plus déterminant. Ce n'est pas que les autres critères soient secondaires — les erreurs ASR sur le jargon ou l'uniformisation vocale sont des défaillances graves —, mais c'est la dimension prosodique qui est la plus irremplaçable. Le contenu gaming, et en particulier le commentaire de partie en direct, vend une présence, une réactivité, une personnalité vocale. Maya, xQc, Tyler1 ou Captain Collins sont reconnaissables à leur façon de parler, de s'exclamer, de rire. Neutraliser cette dimension, c'est neutraliser ce que le spectateur est venu consommer. La relation locuteur-spectateur occupe une place presque aussi centrale, pour la même raison : dans ce genre, la voix est le produit. La traduction automatique et la crédibilité discursive sont des critères importants, mais leurs défaillances produisent des effets locaux (une réplique incompréhensible, un terme mal traduit) là où la neutralisation prosodique produit un effet global qui affecte l'ensemble de la vidéo.

Dans la vulgarisation scientifique, c'est la crédibilité discursive qui occupe la première place. Un exposé scientifique repose sur un contrat de précision entre l'expert et son audience : les termes employés, les noms cités et les chiffres annoncés doivent être exacts, car c'est sur cette exactitude que repose l'autorité du locuteur. Quand le doublage automatique restitue « 8 mètres par seconde » là où l'orateur énonce une valeur exprimée en miles, ou prononce « Liasse » pour Laplace, il ne produit pas une erreur locale : il compromet la crédibilité de l'exposé dans son ensemble. La prosodie et l'expressivité arrivent au second rang, non parce qu'elles importent moins dans l'absolu, mais parce que le discours scientifique survit partiellement à leur neutralisation : un spectateur qui reçoit les arguments de Brian Cox dans un registre plat peut encore les comprendre, même s'il n'en perçoit pas la force rhétorique. La synchronisation occupe en revanche une place plus marginale dans ce sous-corpus que dans le gaming, en

particulier lorsque le locuteur n'est pas visible à l'écran, comme dans plusieurs passages de la vidéo de Cox.

Cette hiérarchisation différenciée confirme ce que Sánchez-Gijón et Kenny (2022, pp. 88-89) posent comme condition de qualité pour tout système de traduction automatique : son adéquation à la fonction communicative du texte source, à son registre et aux conventions du genre. Une traduction qui préserve le contenu sémantique en neutralisant la dimension prosodique peut convenir à un texte à visée informative pure ; elle ne convient pas à un texte dont la fonction est d'incarner une présence, de créer une relation affective ou de démontrer une autorité. Le doublage automatique de YouTube produit, dans les deux genres, un transfert orienté vers une seule fonction — la transmission informationnelle — quelle que soit la fonction réelle du texte source. C'est cette inadéquation entre la logique du système et la diversité des fonctions des contenus auxquels il s'applique qui constitue la limite la plus fondamentale du dispositif. Un dernier élément doit toutefois être versé au dossier. Baños-Piñero et Chaume établissent que, dans le doublage professionnel lui-même, les porteurs d'oralité se concentrent au niveau lexical et sémantique tandis qu'ils demeurent limités au niveau phonétique et prosodique (Baños-Piñero et Chaume 2009, section 4). La défaillance prosodique du doublage automatique prolonge donc une faiblesse ancienne de la pratique plutôt qu'elle n'en révèle une inédite. Ce qui lui est propre, c'est d'échouer simultanément au niveau lexical, là où le doublage humain concentre au contraire l'essentiel de son effort.

11. Conclusion générale

Ce travail est parti d'une question simple dans sa formulation et incertaine dans sa réponse : le doublage automatique de YouTube parvient-il à restituer l'acte communicationnel propre aux contenus de créateurs ? L'analyse de vingt et un extraits répartis sur cinq vidéos permet à présent d'y répondre, et de préciser ce que cette réponse engage sur le plan conceptuel.

11.1 Synthèse des résultats

11.1.1 Par critère

La synthèse par critère peut être conduite brièvement, les chapitres 8 à 10 en ayant établi le détail. L'isochronie est globalement respectée et la fidélité sémantique satisfaisante sur les segments explicatifs simples ; la synchronisation labiale n'est jamais atteinte et n'est pas visée ; la synchronisation cinésique échoue chaque fois que la voix neutralise une emphase que le geste continue de porter. La prosodie constitue le point de défaillance le plus constant, sur la totalité des extraits et indépendamment du genre, du format et de la qualité de la transcription. La crédibilité discursive est atteinte de deux façons distinctes : par la réduction systématique des marqueurs d'oralité dans les deux sous-corpus, et par des erreurs d'exactitude dans le seul sous-corpus scientifique. La traduction automatique échoue à des maillons différents selon le genre — reconnaissance vocale dans le gaming, synthèse vocale dans la vulgarisation —, avec une catégorie d'erreur particulièrement coûteuse, celle qui produit un énoncé vraisemblable mais faux. La relation locuteur-spectateur est enfin le critère le plus uniformément compromis : aucune voix stable n'est attribuée aux locuteurs, et la voix originale demeure audible sous le doublage.

11.1.2 Par genre

Le gaming présente une gradation nette, indexée sur deux variables. La première est la densité de l'expressivité orale : le nombre et l'intensité des variations émotionnelles portées par la voix — rires, cris, accélérations, ruptures de ton. La seconde est la complexité de la configuration sonore : le nombre de locuteurs, la fréquence des chevauchements de parole et le niveau du fond sonore concurrent. Les trois vidéos du sous-corpus se laissent ordonner selon ces deux axes, et les résultats se dégradent à mesure que l'on progresse sur l'un comme sur l'autre. Le tutoriel de Captain Collins, avec un locuteur unique et un registre planifié, fournit les résultats

les moins dégradés ; les segments explicatifs simples sont correctement transcrits et traduits, et le doublage remplit partiellement sa fonction informative. Le stream de Maya introduit la pluralité de locuteurs, le jargon et la réaction en temps réel, ce qui multiplie les occasions d’erreur et compromet la dimension relationnelle. La séquence de Doaenel, avec cinq locuteurs simultanés et une terminologie pour laquelle le système ne dispose d’aucune ressource, produit un résultat qui ne remplit plus aucune des fonctions communicatives du discours original.

La vulgarisation scientifique ne connaît pas de dégradation comparable. Même sur les extraits les plus défaillants, la structure informative demeure décodable : le spectateur francophone suit l’argument d’Andreas Albrecht et comprend l’explication de Brian Cox, même dans un registre plat et avec une terminologie approximative. Le sous-corpus livre par ailleurs le meilleur résultat de l’ensemble analysé, l’extrait 5 du sous-corpus scientifique, dont la réussite relative tient à la conjonction de conditions favorables — locuteur unique, débit posé, syntaxe simple, registre prévisible — et qui délimite par contraste le domaine de fonctionnement acceptable du dispositif : un discours normalisé, continu et monologique.

La comparaison fait ainsi apparaître que les deux genres n’exposent pas les mêmes vulnérabilités. Le gaming repose sur des pratiques communicationnelles codifiées — interpellations de l’audience, remerciements nominatifs, vocabulaire communautaire — que le système ne reconnaît pas et neutralise, atteignant le lien de confiance qui unit le créateur à sa communauté. La vulgarisation repose sur des conventions d’autorité épistémique — exactitude terminologique, attribution des idées, précision des données — que le système compromet par des substitutions et des erreurs factuelles, atteignant la crédibilité intellectuelle de l’exposé. Dans les deux cas, la défaillance touche ce qui fonde le pacte implicite entre l’énonciateur et son public.

11.2 Retour sur les hypothèses

La première hypothèse est confirmée, et la manière dont elle l’est mérite d’être précisée. Confronté aux six standards de qualité par lesquels Chaume définit le doublage, le dispositif en satisfait un, en satisfait partiellement un deuxième, et échoue aux quatre autres. Celui qu’il satisfait est la fidélité de la traduction, globalement correcte sur les segments explicatifs simples. Celui qu’il satisfait partiellement est la synchronisation acceptable, dont Chaume rappelle qu’elle recouvre trois composantes : l’isochronie est généralement respectée, tandis que la synchronisation labiale n’est pas même recherchée et que la synchronisation cinésique

échoue chaque fois que la voix neutralise ce que le geste continue de porter. Les quatre auxquels il échoue sont la crédibilité et le réalisme des dialogues, au sens où l'entendent Marzà et Chaume, pour qui l'essentiel réside dans l'impression produite et la crédibilité de la parole saisie comme un tout (2009, p. 35) ; la cohérence entre image et paroles, rompue chaque fois que la voix neutralise ce que l'image continue de montrer ; la qualité sonore, compromise par la coexistence de deux couches audio dans deux langues ; et le jeu d'acteur, absent par construction, et s'y ajoute une infraction à une norme que Chaume rapporte d'après Díaz Cintas, selon laquelle les mêmes voix ne doivent pas être réutilisées à travers une production (Díaz Cintas 2003, pp. 55-58, cité dans Chaume 2012, p. 20).

Cette répartition n'est pas arbitraire. Ce que le dispositif satisfait est ce dont le respect peut être vérifié sur le texte seul : pour la fidélité, il suffit de comparer le sens de la phrase source et celui de la phrase cible ; pour l'isochronie, seule composante réussie de la synchronisation, de comparer deux durées. Ce sont des opérations qu'une machine sait faire, parce qu'elles ne demandent de considérer ni qui parle ni comment. Les quatre standards auxquels il échoue, ainsi que les deux autres composantes de la synchronisation, exigent au contraire une voix qui interprète et un corps auquel cette voix puisse être rapportée : synchroniser des lèvres suppose de voir un visage, produire un dialogue crédible suppose de savoir à qui l'on parle, accorder le son à l'image suppose de comprendre ce que l'image montre, et jouer suppose un interprète. Le partage entre ce que le système réussit et ce qu'il manque recoupe donc exactement la distinction posée au chapitre 2 entre le contenu propositionnel d'un énoncé et l'acte communicationnel qu'il accomplit.

Il importe de préciser la nature de ces défaillances. Une partie d'entre elles sont contingentes et susceptibles d'être corrigées : les erreurs de reconnaissance portant sur le jargon relèvent d'un déficit de données d'entraînement, les prononciations fautives de noms propres relèvent d'un défaut de ressources lexicales, l'absence de commutation phonologique entre français et anglais relève d'une limite d'implémentation. D'autres sont architecturales : l'absence de module d'identification des locuteurs, l'absence d'incarnation vocale, l'absence de toute étape de révision humaine. Aucune amélioration des modèles ne les résoudra, car elles ne procèdent pas d'une performance insuffisante mais de la conception du dispositif. Cette distinction constitue le principal apport descriptif de l'analyse.

La deuxième hypothèse est confirmée, mais son domaine de validité doit être restreint. La gradation observée dans le sous-corpus gaming la vérifie pleinement : les défaillances

s'aggravent à mesure que croissent la densité de l'expressivité orale et la complexité de la configuration sonore. Elle ne rend cependant pas compte de l'ensemble des résultats. Les erreurs les plus caractéristiques du sous-corpus scientifique surviennent sur la vidéo la plus planifiée du corpus, et portent sur des termes correctement transcrits et traduits. Elles ne dépendent donc pas de la densité de l'oralité, mais d'une variable distincte, que l'on peut décrire comme l'extériorité lexicale : noms propres, mots étrangers, homographes, c'est-à-dire tout élément excédant le français standard.

L'hypothèse gouverne ainsi un segment de la chaîne — la reconnaissance vocale, et par répercussion la traduction — sans gouverner la synthèse vocale, qui obéit à une autre logique de défaillance. Cette restriction n'affaiblit pas le résultat ; elle le précise. Elle éclaire aussi l'observation de Baños-Piñero et Chaume selon laquelle les porteurs d'oralité du doublage professionnel se concentrent au niveau lexical et sémantique, demeurent limités au niveau phonétique et prosodique et sont pratiquement inexistantes au niveau morphologique (Baños-Piñero et Chaume 2009, section 4). Le doublage automatique échoue au niveau prosodique, là où le doublage humain était déjà le moins investi, mais il échoue également au niveau lexical, là où celui-ci concentre l'essentiel de son effort. Sa défaillance prolonge une faiblesse ancienne de la pratique et lui en ajoute une nouvelle.

La troisième hypothèse est confirmée dans son intuition et modifiée dans sa formulation. Elle posait que le dispositif se rapproche davantage du voice-over que du doublage. L'analyse conduit à une conclusion plus radicale : il ne relève pleinement d'aucun des deux modes, et c'est cette position qui justifie de le nommer.

11.3 Le voice-over algorithmique

Trois considérations soutiennent la proposition conceptuelle formulée ici.

La première est architecturale, et le chapitre 3 en a posé les termes : la chaîne mise en œuvre par YouTube reprend celle que le projet ALST avait modélisée pour le voice-over, mais en supprime la révision humaine que ce modèle plaçait après chacune de ses trois étapes. Ce qui se joue dans cette suppression peut maintenant être nommé : l'automatisation y cesse d'être une assistance au traducteur pour devenir une substitution, et c'est cette substitution — non l'automatisation en elle-même — qui produit les défaillances architecturales relevées au chapitre 10.

Le résultat d'Ortiz-Boix et Matamala rappelé au chapitre 3 — l'écart entre traduction humaine et post-édition n'apparaît qu'une fois le texte mis en voix — trouve ici sa confirmation dans un contexte plus défavorable encore, puisque le dispositif de YouTube ne comporte, lui, aucune post-édition. Sur l'ensemble du corpus, la fidélité du transfert textuel coexiste régulièrement avec l'échec du produit vocal, et c'est cette dissociation entre le texte et sa réalisation sonore qui constitue la signature du dispositif.

La deuxième considération est fonctionnelle. Le rapprochement avec le voice-over s'impose au premier examen. La voix originale demeure audible sous la voix traduite ; le transfert privilégie l'information sur la recreation ; les marqueurs d'oralité spontanée disparaissent. Ces trois traits correspondent à ceux que la recherche attribue au voice-over : Franco, Matamala et Orero observent que les hésitations, répétitions, faux départs et anomalies syntaxiques disparaissent de la version *voiced-over* pour atteindre l'isochronie et privilégier la compréhension, et que, là où le doublage s'efforce de recréer des dialogues crédibles, le voice-over privilégie la transmission intelligible d'un contenu (2010, p. 74, cité dans Matamala 2018, p. 72).

L'examen des conventions du voice-over interdit toutefois d'en rester là. Ce mode impose quatre synchronies : l'isochronie de voice-over, par laquelle la traduction commence quelques mots après l'original et s'achève avant lui ; la synchronie littérale, qui veut que les mots audibles en début et en fin d'unité, non recouverts, soient traduits littéralement afin que les spectateurs comprenant la langue source puissent rapporter la traduction à l'original ; la synchronisation cinésique ; la synchronie d'action (Matamala 2020, pp. 134-135). Le corpus analysé documente des défaillances sur les deux dernières. Quant à la présence audible de l'original, elle obéit dans le voice-over à une enveloppe de volume réglée : l'original est perceptible en clair à l'attaque, atténué pendant la traduction, puis ramené à son niveau normal lorsque celle-ci s'achève (Díaz Cintas et Orero 2010, p. 441). Cette organisation remplit une fonction que Matamala décrit comme une illusion d'authenticité, renforcée lorsque la synchronie littérale est respectée (2018, p. 72).

Le dispositif de YouTube ne reproduit rien de cette organisation. L'atténuation y est plate et continue, sans attaque en clair, sans reprise finale, sans traduction littérale des segments non recouverts, et le signal source perce le filtre sur les pics d'intensité. Le même phénomène acoustique remplit donc une fonction inverse : ce qui authentifie le voice-over expose ici l'appareil. Il en va de même de l'effacement des marqueurs d'oralité, qui procède dans le voice-

over d'un choix éditorial motivé et réversible, et résulte ici du traitement de ces marqueurs comme perturbations du signal par le module de reconnaissance. Le résultat coïncide ; le statut diffère entièrement.

Cette inversion des fonctions se laisse formuler plus précisément encore à partir de l'opposition posée au chapitre 1. Díaz Cintas et Orero distinguent les deux modes par le régime de visibilité qu'ils imposent à la traduction : le doublage la dissimule, le voice-over la rappelle (2010, p. 442). Or le dispositif étudié occupe les deux positions à la fois. Il se présente comme un doublage, et prétend donc faire oublier au spectateur qu'une traduction a eu lieu ; mais il laisse simultanément la voix source audible sous la voix synthétique, et rappelle ainsi, à chaque instant où le filtre d'atténuation cède, qu'une opération de transfert est en cours. Il revendique le régime du doublage tout en produisant les effets du voice-over.

Il faut se garder d'y voir un simple défaut d'exécution, qu'un meilleur traitement du signal viendrait corriger. La coexistence des deux régimes procède de ce que le dispositif n'a pas choisi le sien. Le doublage dissimule la traduction parce qu'il en assume le coût : il recrute des comédiens, ajuste la durée des énoncés, apparie une voix à un corps. Le voice-over rappelle la traduction parce qu'il en fait une convention : il règle l'enveloppe sonore, traduit littéralement les segments non recouverts, laisse au spectateur les moyens de rapporter ce qu'il entend à ce qui a été dit. Le doublage automatique de YouTube ne consent ni à l'un ni à l'autre. Il hérite de la dissimulation son ambition — se donner pour la parole même du créateur — et du rappel sa manifestation — la persistance de la voix originale —, sans les moyens qui rendent l'une crédible et l'autre lisible.

La troisième considération est phénoménologique. Bosseaux soutient que le doublage produit un effet d'inquiétante étrangeté, parce que le spectateur entend une voix qui n'appartient pas au corps qu'il voit jouer, et propose de qualifier la voix doublée d'acousmatique, au sens que Chion (1982) donne à ce terme (Bosseaux 2015, p. 213). Elle mobilise à cet égard la description que Dolar donne de cette voix : « une voix en quête d'une origine, en quête d'un corps, mais qui, même lorsqu'elle trouve son corps, révèle que cela ne fonctionne pas tout à fait » (2006, p. 61, cité dans Bosseaux 2015, p. 214). La voix doublée peut dès lors être décrite comme un fantôme hantant le corps d'un autre acteur.

La transposition à la voix de synthèse déplace le problème. Chez Bosseaux, la disjonction s'établit entre deux corps, celui que l'on voit et celui, invisible, du comédien de doublage. Une

voix de synthèse n'a pas de corps d'origine. Elle n'est pas en quête d'un corps au sens de Dolar, n'en ayant jamais eu. Le grain, que Barthes définit comme ce qui parvient à l'oreille depuis les cavités, les muscles, les membranes et les cartilages (1977, p. 179, cité dans Bosseaux 2015, p. 51), n'y est pas atténué mais structurellement absent. La métaphore du fantôme y trouve sa limite, puisqu'elle suppose un corps quitté.

On propose en conséquence d'appeler *voice-over algorithmique* un mode de revoicing entièrement automatisé, caractérisé par la suppression de toute étape de révision humaine dans une chaîne conçue pour en comporter ; par la production d'un substitut vocal standardisé orienté vers le seul transfert informatif, quelle que soit la fonction communicationnelle du texte source ; par la reproduction des traits de surface du voice-over accompagnée de l'inversion de leurs fonctions ; et par une voix acousmatique sans origine, qui ne peut satisfaire à la condition d'incarnation définissant le doublage.

L'intérêt de cette notion n'est pas de disqualifier le dispositif, mais de le décrire correctement. Évalué comme un doublage, il apparaît comme un doublage raté, ce qui invite à en attendre l'amélioration ; décrit comme un mode distinct, il apparaît comme un dispositif accomplissant exactement ce pour quoi il est conçu, dont certaines limites sont contingentes et d'autres constitutives. La désignation retenue par la plateforme induit ainsi une erreur de catégorie que ce travail se propose de corriger. Díaz Cintas formule à ce sujet une remarque qui éclaire la situation. Un texte audiovisuel combine des images, des sons, de la musique, du texte écrit et de la parole, mais seule cette dernière est réellement transformée lors d'une traduction ; tout le reste est censé traverser les frontières culturelles sans perte, comme s'il s'agissait d'un espéranto que tout spectateur comprendrait de la même façon. Il tient cette supposition pour manifestement fautive (Bogucki et Díaz Cintas 2020). Le doublage automatique la pousse à son terme : en traitant le seul flux verbal et en ignorant l'image, il agit exactement comme si le reste du texte audiovisuel n'avait pas besoin d'être traduit.

11.4 Limites de l'analyse

Ce travail comporte des limites dont plusieurs conditionnent la portée de ses conclusions.

La première tient au corpus. Vingt et un extraits issus de cinq vidéos, dans une seule paire de langues et sur deux genres, ne permettent pas de généraliser à l'ensemble des contenus concernés. Le choix de l'anglais vers le français porte de surcroît sur une combinaison

favorable, abondamment représentée dans les données d'entraînement des systèmes mobilisés ; les résultats obtenus constituent vraisemblablement une borne haute.

La deuxième tient à l'absence d'étude de réception. L'argumentation développée ici repose largement sur ce que le spectateur est supposé percevoir : l'incohérence entre un corps et une voix, la platitude d'une intonation, la rupture d'un contrat de crédibilité. Ces effets sont inférés de l'analyse, non mesurés auprès d'un public. Cette limite est d'autant plus sensible que l'analyste est un traducteur en formation, dont l'attention aux écarts diffère de celle d'un spectateur ordinaire, lequel pourrait fort bien accepter le dispositif pour ce qu'il permet plutôt que le juger sur ce qu'il manque.

La troisième tient à l'instabilité de l'objet. Les modèles qui alimentent le système sont régulièrement mis à jour, de sorte que les observations rapportées correspondent à l'état du dispositif au moment du visionnage. La reproductibilité exacte de l'analyse ne peut être garantie, et certaines défaillances identifiées comme contingentes auront peut-être disparu à la lecture de ce travail — ce qui, au demeurant, constituerait une confirmation de la distinction proposée entre défaillances contingentes et défaillances architecturales.

La quatrième tient à la part de subjectivité inhérente à la sélection des extraits. Le protocole de repérage explicité au chapitre 6 vise à la contenir, mais le choix des passages retenus procède d'un jugement de pertinence qui ne peut être entièrement formalisé.

11.5 Perspectives

Plusieurs prolongements se dégagent de ce travail.

Le plus nécessaire concerne la réception. Matamala relève à deux reprises le manque d'études de réception sur le voice-over, notamment sur l'effet d'une voix unique lisant plusieurs locuteurs, Bogucki observant que les dialogues impliquant plusieurs locuteurs sont difficiles à suivre lorsqu'ils sont lus par un seul lecteur (2013, p. 6, cité dans Matamala 2018, p. 74). La question se pose avec une acuité particulière pour le doublage automatique, dont le public n'est pas un public de doublage constitué mais l'audience mondiale d'une plateforme, dont on ignore les attentes et les seuils de tolérance. Bosseaux formule d'ailleurs le même constat pour le doublage professionnel : davantage d'études de réception seraient nécessaires pour établir quel est l'effet réel de l'attribution des voix sur les produits doublés (2018, p. 54). Un protocole

comparant les jugements de spectateurs exposés à la version originale sous-titrée, au doublage automatique et à un doublage professionnel permettrait de tester empiriquement les inférences avancées ici.

Un deuxième prolongement concerne l'extension du corpus : à d'autres paires de langues, en particulier vers des langues typologiquement éloignées de l'anglais ou moins dotées en ressources ; à des pays de tradition non doublante, où les attentes du public à l'égard du revoicing diffèrent ; et à d'autres genres, notamment les contenus à forte charge affective ou les formats brefs.

Un troisième prolongement est diachronique. Le suivi du même corpus à intervalles réguliers permettrait de mesurer lesquelles des défaillances relevées se résorbent et lesquelles persistent, et de mettre à l'épreuve la distinction entre limites contingentes et limites architecturales sur laquelle repose une partie de l'argumentation.

Un quatrième prolongement touche à la position des créateurs. Le dispositif s'applique par défaut, mais peut être désactivé. Les motifs de ce choix, les stratégies adoptées par ceux qui refusent le doublage automatique, et l'éventuelle transformation de leurs pratiques d'écriture en anticipation d'un traitement automatique constituent un objet distinct, relevant de la sociologie des pratiques autant que de la traductologie. Miggiani (2025) a ouvert une voie comparable du côté des professionnels du doublage.

Un dernier prolongement, plus prospectif, concerne l'évolution technique elle-même. Bywood signale que les technologies de manipulation vidéo permettant de faire prononcer à un individu des paroles qu'il n'a pas tenues auraient des applications évidentes dans le doublage interlinguistique si elles devenaient accessibles à un coût raisonnable (2020, p. 513). Une telle évolution modifierait les termes du problème posé ici. La question de l'incarnation vocale ne se poserait plus dans les mêmes termes si le corps visible à l'écran était lui-même synthétisé pour s'accorder à la voix entendue : la cohérence entre corps et voix serait alors rétablie par le bas, non par la restitution d'une présence mais par la fabrication conjointe des deux termes. Ce que ce travail décrit comme une voix sans corps pourrait céder la place à une configuration où ni la voix ni le corps ne renverraient plus à une personne. Il reviendrait alors à la discipline de déterminer si une telle production relève encore de la traduction.

Bibliographie

- Austin, John L. *How to Do Things with Words*. Oxford University Press, 1962.
- Baños-Piñero, Rocío, et Frederic Chaume. « Prefabricated Orality : A Challenge in Audiovisual Translation. » *inTRAlinea*, numéro spécial « The Translation of Dialects in Multimedia », 2009, www.intralinea.org/specials/article/1714.
- Barthes, Roland. *Image, Music, Text*. Traduit par Stephen Heath, Fontana, 1977.
- Bogucki, Łukasz. *Areas and Methods of Audiovisual Translation Research*. Peter Lang, 2013.
- Bogucki, Łukasz, et Jorge Díaz Cintas. « An Excursus on Audiovisual Translation. » *The Palgrave Handbook of Audiovisual Translation and Media Accessibility*, dirigé par Łukasz Bogucki et Mikołaj Deckert, Palgrave Macmillan, 2020, pp. 11-32.
- Bosseaux, Charlotte. *Dubbing, Film and Performance : Uncanny Encounters*. Peter Lang, 2015.
- Bosseaux, Charlotte. « Investigating Dubbing : Learning from the Past, Looking to the Future. » *The Routledge Handbook of Audiovisual Translation*, dirigé par Luis Pérez-González, Routledge, 2018, pp. 48-63.
- Bywood, Lindsay. « Technology and Audiovisual Translation. » *The Palgrave Handbook of Audiovisual Translation and Media Accessibility*, dirigé par Łukasz Bogucki et Mikołaj Deckert, Palgrave Macmillan, 2020, pp. 503-517.
- Caillé, Pierre-François. « Cinéma et traduction : le traducteur devant l'écran. » *Babel*, vol. 6, n° 3, 1960, pp. 103-109.
- Chaume, Frederic. *Audiovisual Translation : Dubbing*. Routledge, 2012.
- Chaume, Frederic. « Dubbing. » *The Palgrave Handbook of Audiovisual Translation and Media Accessibility*, dirigé par Łukasz Bogucki et Mikołaj Deckert, Palgrave Macmillan, 2020, pp. 103-132.
- Chion, Michel. *La Voix au cinéma*. Cahiers du cinéma / Éditions de l'Étoile, 1982.

- Chion, Michel. *Le Son au cinéma*. Cahiers du cinéma, 1985.
- Darwish, Ali, et Pilar Orero. « Rhetorical Dissonance of Unsynchronized Voices : Issues of Voice-Over in News Broadcasts. » *Babel*, vol. 60, n° 2, 2014, pp. 129-144.
- Díaz Cintas, Jorge, et Lydia Hayes. « Role Reversal : An Overview of Audiovisual Translation into English. » *Íkala, Revista de Lenguaje y Cultura*, vol. 28, n° 2, 2023, pp. 1-18.
- Díaz Cintas, Jorge, et Pilar Orero. « Voiceover and Dubbing. » *Handbook of Translation Studies*, vol. 1, dirigé par Yves Gambier et Luc van Doorslaer, John Benjamins, 2010, pp. 441-445.
- Dolar, Mladen. *A Voice and Nothing More*. MIT Press, 2006.
- Díaz Cintas, Jorge. *Teoría y práctica de la subtitulación : inglés-español*. Ariel, 2003.
- Franco, Eliana, Anna Matamala et Pilar Orero. *Voice-over Translation : An Overview*. Peter Lang, 2010.
- Gregory, Michael, et Susanne Carroll. *Language and Situation : Language Varieties and Their Social Contexts*. Routledge & Kegan Paul, 1978.
- Kenny, Dorothy, dir. *Machine Translation for Everyone : Empowering Users in the Age of Artificial Intelligence*. Language Science Press, 2022.
- Lotman, Iouri. *La Structure du texte artistique*. Traduit par Anne Fournier et al., Gallimard, 1973 [éd. russe 1970].
- Marzà, Anna, et Frederic Chaume. « The Language of Dubbing : Present Facts and Future Perspectives. » *Analysing Audiovisual Dialogue : Linguistic and Translational Insights*, dirigé par Maria Freddi et Maria Pavesi, CLUEB, 2009, pp. 31-39.
- Matamala, Anna. « The ALST Project : Technologies for Audiovisual Translation. » *Proceedings of the 37th Conference Translating and the Computer*, AsLing, 2015, pp. 79-89.

- Matamala, Anna. « Voice-over : Practice, Research and Future Prospects. » *The Routledge Handbook of Audiovisual Translation*, dirigé par Luis Pérez-González, Routledge, 2018, pp. 64-81.
- Matamala, Anna. « Translating Non-fictional Genres : Voice-over and Off-screen Dubbing. » *The Palgrave Handbook of Audiovisual Translation and Media Accessibility*, dirigé par Łukasz Bogucki et Mikołaj Deckert, Palgrave Macmillan, 2020, pp. 133-148.
- Meyer-Dinkgräfe, Daniel. « Thinking Voice in Dubbing. » *Consciousness, Literature and the Arts*, vol. 8, n° 2, 2007.
- Miggiani, Giselle Spiteri. « Rethinking Creativity in Dubbing : Potential Impact of AI Dubbing Technologies on Creative Practices, Roles and Viewer Perceptions. » *Translation Spaces*, John Benjamins, 2025.
- Mingant, Nolwenn. « Tarantino's *Inglourious Basterds* : A Blueprint for Dubbing Translators ? » *Meta*, vol. 55, no 4, 2010, pp. 712-731.
- Ortiz-Boix, Carla, et Anna Matamala. « Post-editing Wildlife Documentary Films : A New Possible Scenario ? » *The Journal of Specialised Translation*, n° 26, 2016, pp. 187-210.
- Pérez-González, Luis, dir. *The Routledge Handbook of Audiovisual Translation*. Routledge, 2018.
- Pérez-Ortiz, Juan Antonio, Mikel L. Forcada et Felipe Sánchez-Martínez. « How Neural Machine Translation Works. » *Machine Translation for Everyone*, dirigé par Dorothy Kenny, Language Science Press, 2022, pp. 141-164.
- Rehm, Georg, et Hans Uszkoreit, dir. *Strategic Research Agenda for Multilingual Europe 2020*. META-NET, 2012.
- Ruiz, Nicholas, et Marcello Federico. « Assessing the Impact of Speech Recognition Errors on Machine Translation Quality. » *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas (AMTA 2014)*, vol. 1 : MT Researchers, dirigé par Yaser Al-Onaizan et Michel Simard, Vancouver, 2014, pp. 261-274.

- Smith, Susan. « Voices in Film. » *Close-Up 02*, dirigé par John Gibbs et Douglas Pye, Wallflower Press, 2007, pp. 159-238.
- Sánchez-Gijón, Pilar, et Dorothy Kenny. « Selecting and Preparing Texts for Machine Translation : Pre-editing and Writing for a Global Audience. » *Machine Translation for Everyone : Empowering Users in the Age of Artificial Intelligence*, dirigé par Dorothy Kenny, Language Science Press, 2022, pp. 81-104.
- Sánchez Mompeán, Sofía. « Dubbing Attitudes through Tonal Patterns : When Tones Speak Louder than Words. » *VIAL*, n° 13, 2016, pp. 135-156.
- Sanz Ortega, Alejandro. *Multilingualism in Film : Translation and Multimodal Methodology*. Thèse de doctorat, University of Edinburgh, 2015.
- Schröder, Marc. « Expressive Speech Synthesis : Past, Present, and Possible Futures. » *Affective Information Processing*, dirigé par Jianhua Tao et Tieniu Tan, Springer, 2009, pp. 111-126.
- Tomaszkiewicz, Teresa. « Transfert des références culturelles dans les sous-titres filmiques. » *(Multi)Media Translation : Concepts, Practices and Research*, dirigé par Yves Gambier et Henrik Gottlieb, John Benjamins, 2001, pp. 237-247.
- van Leeuwen, Theo. *Speech, Music, Sound*. Macmillan, 1999.
- Vaswani, Ashish, et al. « Attention Is All You Need. » *Advances in Neural Information Processing Systems 30*, 2017, pp. 5998-6008.
- Whitman-Linsen, Candace. *Through the Dubbing Glass : The Synchronization of American Motion Pictures into German, French and Spanish*. Peter Lang, 1992.
- YouTube Support. *Utiliser le doublage automatique*. [support.google.com/youtube/answer/15569972 ?hl=fr](https://support.google.com/youtube/answer/15569972?hl=fr). Consulté le 12 novembre 2025.

Filmographie

Big Think. « Brian Cox on How Black Holes Could Unlock the Mysteries of Our Universe. » *YouTube*, 27 octobre 2023, 12 min 51 s, www.youtube.com/watch?v=pGsbEd6w7PI.

Captain Collins. « Outer Wilds Starter Guide (NO SPOILERS) — How to Play This Game — Out on Steam in 2020. » *YouTube*, 1er juillet 2020, 7 min 10 s, www.youtube.com/watch?v=IbtGdABmVpQ.

Doaenel. « xQc Lost His MIND after Kadeem Said This to Him... ft Tyler1 & Tobias Fate. » *YouTube*, 8 juillet 2026, 1 h 49 min 06 s, www.youtube.com/watch?v=wG1kaXTedLI.

Maya. « Un non-joueur peut-il battre un boss d'Elden Ring ? » *YouTube*, 14 juillet 2025, 32 min 38 s, www.youtube.com/watch?v=vG5Zmjkt48U.

World Science Festival. « Straight Talk on Cosmic Origins. » *YouTube*, 10 janvier 2025, 1 h 34 min 54 s, www.youtube.com/watch?v=Id5iHy59UL8.

Annexes

Les fiches d'analyse complètes des vingt et un extraits figurent au chapitre 8 et au chapitre 9.
Aucune annexe distincte n'est jointe au travail.