ROSETTA
accès multilingue

RObot de Sous-titrage Et Toute Traduction Adaptés

Livrable Résultats de l'évaluation humaine

6.3.4

Sous-titre 1

Durée de projet 42 mois : Octobre 2018 – Novembre 2021

Historique	Date	Modification(s)
V 0.00	16/11/2021	Rédigé par Taina Victor plan provisoire
V 1	25/11/2021	Rédigé par Taina Victor
V2	25/11/2021	Validation version provisoire et relecture Léa Lachaud et Taïna Victor
V3		

Auteurs du livrable LUTIN

- LUTIN USERLAB

- Lachaud Léa
- Victor Taïna
- Tijus Charles

Ce livrable répond à la tâche 6.3

Afin de (descriptif de l'objectif tu livrable)

Mots clés

- Secteur(s) d'application : TBD
- Domaine(s) technologiques : intelligence artificielle, apprentissage profond, Big Data, apprentissage automatique, corpus, génération automatique des sous-titrages adaptés multilingues, génération de contenus en langue des signes, avatar signant, signeur virtuel, animation d'avatar, capture de mouvements

Table des matières

Figures

Tableaux

Résumé

Le livrable « *6.3.4: Résultats de l'évaluation humaine : synthèse* » est une synthèse de l'ensemble des résultats obtenus tout au long du projet pour les évaluations des modules LSF et Sous-titrage de ROSETTA. Il est un résumé des livrables suivants (dont les livrables intermédiaires):

6.3.1 : Résultats de l'évaluation technique 1 : Recommandations pour le sous-titrage et la LSF
6.3.2 : Résultats de l'évaluation technique 2 : questionnaires en ligne
6.3.3 : Résultats de l'évaluation technique 3 : résultats sur sous-titres bruités

Ces résultats ont été obtenus en utilisant les méthodes d'observation, d'expérimentation (en présence et à distance) et d'analyse déterminées dans le lot 6.2.

L'appréciation par les participants du module sous-titre et celle du module LSF ont été observées pour fournir des retours. Les résultats ont concerné les réponses obtenues aux deux échelles ergonomiques élaborées dans le cadre du projet. Ces échelles servent à évaluer 7 composantes ergonomiques qui favorisent l'utilisabilité d'un système. Chaque module a ainsi été évalué en ligne selon ces dimensions.

Les analyses ont révélé une bonne fiabilité de l'échelle de mesure à 13 items. Toutefois, ce nombre a été abaissé à 12 items pour en augmenter les qualités psychométriques.

Un second objectif a été, pour le module LSF, de distinguer les apports de la méthode multicanaux par rapport à d'autres méthodes de générations et également par rapport à un signeur virtuel exemplaire : l'humain.

Si les vidéos du signeur virtuel humain présentent des qualités ergonomiques mieux évaluées comparées à leur génération automatique ; on note que la méthode multicanaux fournit de meilleurs scores. Il est intéressant de noter la nette appréciation des utilisateurs pour la méthode automatique multicanaux privilégiée dans le cadre du projet. ROSETTA (LSF) semble ainsi mieux correspondre aux attentes des utilisateurs ; ROSETTA-LSF pouvant être encore amélioré.

Pour le sous-titrage, les nouvelles versions de ROSETTA aux plus anciennes pour mettre en évidence des signes d'amélioration. Les résultats ont montré que le traditionnel présentait des scores plus élevés que les autres modes de sous-titrage. De plus, les performances de ROSETTA français étaient supérieures à ROSETTA multilingue sauf pour la composante Utilité. Cela peut être dû à l'influence de la manipulation de deux langues qui peut accentuer l'effort cognitif quand on lit les sous-titres anglais de programmes en français. En revanche, par rapport au sous-titrage de YouTube ROSETTA Fr (138), la plus récente ne présente pas de meilleurs scores. Notons que ROSETTA multilingue (anglais) présente des scores supérieurs à Youtube. Toutefois, plus le niveau des participants est faible, plus ils évaluaient les sous-titres anglais de bonne qualité.

Enfin, les résultats utilisant des sous- titrages bruités, montrent que les appréciations recueillies selon le niveau de bruitage peuvent servir de métrique pour évaluer la qualité des sous-titrages, selon la correspondance entre appréciations ; ceci en comparant celles du sous-titrage cible à évaluer avec celles des corpus plus ou moins bruités.

Introduction

1 Contexte

1.1 Objectifs

Ce livrable a pour objectif de présenter les résultats des méthodes mises en œuvre dans la tâche 6.2.

Les résultats concernent la mesure des degrés d'utilité et d'utilisabilité des modules pour chacune des déficiences, à travers les recommandations publiées dans la littérature. Principalement, présenter les évaluations pour chaque public des recommandations existantes.

L'objectif est également de présenter les résultats de la mesure de la compréhension/ la comprenabilité et la satisfaction des utilisateurs en considérant les différentes dimensions de l'usage de ROSETTA, notamment les dimensions et usages : Apprentissage, Information et Divertissement.

Le présent livrable a pour objectif de montrer la pertinence des modules automatiques pour les usagers en fonction du degré de bruit associés aux sous-titres.

1.2 Le contexte

Le présent livrable s'inscrit dans le cadre du lot 6 et de la synthèse des résultats des méthodes élaborées au cours du projet. Ces méthodes ont eu pour objectif d'évaluer le prototype ROSETTA par rapport à l'existant :

- Le sous-titrage traditionnel et automatique
- Les méthodes de génération automatiques de LSF

2 Synthèse résultats évaluation technique – 1 – grilles d'évaluation experte – L 6.3.1

Le livrable 6.3.1 exposait les résultats des trois questionnaires portant sur l'évaluation des recommandations : questionnaire de recommandations LSF, questionnaire de recommandations de sous-titrage classique (utilisateurs), et questionnaire de recommandations de sous-titrage expert (professionnels) (cf. méthode livrable 6.2.2).

L'objectif de cette étude était la constitution de 3 grilles d'évaluation expertes permettant de prioriser les recommandations en fonction de leurs importances. L'importance de chaque item du questionnaire (correspondant à une recommandation) était évaluée par les différents groupes interrogés (utilisateurs de sous-titrage, professionnels du sous-titrage, et utilisateurs de la LSF). Par la suite, les résultats ont été hiérarchisés dans des grilles d'évaluation expertes, et un coefficient de pondération était attribué à chaque item de la grille. L'objectif de ces grilles d'évaluation expertes étaient d'attribuer un score d'exemplarité ergonomique aux modules évalués (vidéos comportant du sous-titrage et de la LSF).

2.1 Cohérence des questionnaires

2.1.1 Questionnaire de recommandations techniques

Le questionnaire de recommandation technique était composé de 18 items classés selon les trois dimensions suivantes :

- La qualité de la langue (6 items)
- La qualité de l'affichage des sous-titres (7 items)
- Le rythme des sous-titres (5 items)

2.1.1.1 Participants

Au total, 26 personnes ont répondu au questionnaire (20 femmes, 3 hommes, 3 non-précisé). La moyenne d'âge était de 47 ans (*ET* = 14,15). 13 participants étaient en freelance et 12 étaient salariés dans le secteur public/privé. Ils avaient en moyenne 14,29 ans (*ET* = 9,42) d'expérience professionnelle dans le domaine du sous-titrage. Les participants sous-titraient essentiellement du français.

2.1.1.2 Mesure de la cohérence interne du questionnaire

Le calcul du coefficient alpha de Cronbach montre une forte cohérence du questionnaire technique (*α de Cronbach* = 0.81).

En revanche, le calcul du coefficient alpha de Cronbach ne montre pas de cohérence pour les dimensions qualité de langue et qualité de l'affichage des sous-titres, avec respectivement un *α de Cronbach* = 0.48 et 0.54). On observe une faible cohérence de la dimensions rythme des sous-titres, avec un *α de Cronbach* = 0.66).

2.1.2 Questionnaire de recommandations classiques

Le questionnaire de recommandations classiques était composé de 28 items classés selon les quatre dimensions suivantes :

- L'affichage contextuel (7 items)
- La qualité de la langue (4 items)
- L'ergonomie du sous-titre (13 items)
- L'accessibilité de l'interface (4 items)

2.1.2.1 Participants

Au total, 142 personnes de 47 ans (*ET*=20) en moyenne ont répondu au questionnaire (94 femmes (*M*=50 ; *ET* =19), 48 hommes (*M*=40 ; *ET* = 20). Les participants ont été répartis en 3 classes d'âge : de 18 à 35 (n=55), de 36 à 65 ans (n=54) et les +65 ans (n=33).

- Chez les moins de 35 ans ,82% ont un niveau d'étude supérieur ou égal au BAC.
- Chez 36- 65 ans, 41% ont au moins un niveau Master
- Chez les +65 ans environ 66% ont un niveau Licence ou plus

Le niveau d'étude de l'ensemble de l'échantillon était élevé

Des analyses complémentaires avec le test non-paramétrique de Mann-Whitney ont mis en évidence des différences d'usages pour la langue de sous-titrage entre les groupes. Les adultes regardaient plus de sous-titres en français (*M*=0.90 *ET*=0.29) que les jeunes (*M*=0.65 *ET*=0.48), *W*=1860,5 ; p=0.002 . Les Jeunes adultes regardaient plus de sous titres en anglais (*M*= 0.5; *ET*=0.5) que les adultes (*M*=0.24 ; *ET*=0.432) et les seniors (*M*=0.03 ; *ET*=0.17), W=1086.5 ,p=0.004 et W=1342 , *p*<0.01 .

Les +65 ans ont affirmé regarder plus de films (*M*=4.62; *ET*=1.34) comparés aux Jeunes adultes (*M*= 3.48 ; ET= 1.2) et aux adultes(M=4.04; *ET*=1.73), avec respectivement *W*=1836 ; *p*= .002 et *W*= 617; *p*=.021 .

Les +35 ans ont rapporté regarder plus les journaux télévisés, les jeux, les documentaires et les talkshows, comparé aux -de 35 ans. Aucune différence entre les +65 et les + de 35 ans sur ces types de programmes n'a été établie.

Les -35 ans ont déclaré utiliser davantage le téléphone comme support (*M*=3.36; ET=1.7) comparé aux + de 35 ans (*M*=2.73 ET=1.44), *W*= 1030 ; *p*=.022.

2.1.2.2 Mesure de la cohérence interne du questionnaire

Le calcul du coefficient alpha de Cronbach montre une forte cohérence du questionnaire classique pour l'échantillon global (*α de Cronbach* = 0.88).

En revanche, le calcul du coefficient alpha de Cronbach ne montre pas de cohérence pour les dimensions qualité de langue et accessibilité de l'interface, avec respectivement un *α de Cronbach* = 0.35 et 0.56). On observe une bonne cohérence des dimensions affichage contextuel et ergonomie du sous-titre rythme des sous-titres, avec respectivement un α *de Cronbach* = 0.71 et 0.78).

L'alpha de Cronbach a également été calculé en fonction des groupes (- 35 ans, 36-65 ans et + 65 ans). Les données sont disponibles en Annexe 1 et dans le livrable 6.3.1.

1.1.1 Questionnaire de recommandations LSF

Le questionnaire de recommandations LSF était composé de 26 items classés selon les quatre dimensions suivantes :

- La qualité de la langue des signes (3 items)
- La qualité d'Interprétation (5 items)
- La qualité de présentation LSF (14 items)
- L'action de l'utilisateur (4 items)

2.1.2.3 Participants

Au total, 30 personnes ont répondu au questionnaire (23 femmes et 7 hommes). La moyenne d'âge était de 43 ans (*ET* = 12). 18 participants étaient sourds ,11 participants étaient malentendants et 1 était entendant.

Parmi les 30 répondants, la moitié avait un niveau de LSF classé C2. La seconde moitié était répartie de la façon suivante : 7 participants avaient un niveau C1, 4 avaient un niveau B2, 3 avaient un niveau B1, et 1 avait un niveau A2.

Plus de la moitié de l'échantillon, soit 18 personnes, se déclaraient bilingues LSF et langue française. 11 participants déclaraient le français comme étant leur langue maternelle et 1 participant déclarait avoir un niveau de langue française intermédiaire.

La majorité des participants(n=12), déclarait avoir un niveau d'étude de master. 5 participants avaient un niveau licence, 5 participants avaient un niveau BTS/DUT, 4 participants avaient un niveau Baccalauréat, 2 participants avaient un niveau de Doctorat, et 2 participants avaient un niveau inférieur au Baccalauréat.

Plus de la moitié de l'échantillon, soit 16 participants, déclarait avoir grandi en utilisant et en côtoyant le milieu de la LSF. 11 participants déclaraient avoir des parents ou des proches qui utilisaient la LSF à la maison, et 13 participants déclaraient avoir utilisé la LSF à l'école.

2.1.2.4 Mesure de la cohérence interne du questionnaire

Le questionnaire LSF ne présente pas de cohérence des items (alpha inférieur à 0.65). On observe seulement une tendance de la dimension qualité de l'interprétation.

2.2 Pondération de la grille d'évaluation experte et comparaisons des dimensions

Le coefficient de pondération correspondait à la moyenne calculée pour chaque item. Les recommandations considérées comme les plus importantes avaient une moyenne supérieure ou égale à la moyenne générale de chaque questionnaire. Des comparaisons entre les dimensions des questionnaires a également été effectuée.

2.2.1 Questionnaire de recommandations techniques

Les recommandations ont été triées par ordre d'importance, c'est-à-dire, du coefficient le plus élevé au plus faible. La moyenne globale du questionnaire technique était évaluée à 5.51 (*ET* = 1.03) (cf. 2.3. Grilles constituées par module et par groupe).

Des différences entre les dimensions du questionnaire ont également été constatées. Le test non-paramétrique de Wilcoxon a mis en évidence des différences entre les dimensions ($p < .05$). La dimension **qualité de la langue** était jugée plus importante (*M*=5.68; *ET*=0.83) que la dimension qualité de l'**affichage des sous-titres** (*M*=5.34 ; *ET*=1.45) , W=278; *p*=.001. La dimension **qualité de la langue** était jugée moins importante (*M*=5.68; *ET*=0.83) que la dimension **rythme des sous-titres** (*M*=5.84 ; *ET*=0.44) , W=85; *p*=.04. La dimension **qualité de l'affichage des sous-titres** est jugée moins importante (*M*=5.34; *ET*=1.453) que la dimension **rythme des sous-titres** (*M*=5.84 ; *ET*=0.44) , W=16; *p*<.001.

Pour les participants, la dimension rythme des sous-titres était plus importante que les deux autres dimensions. La seconde dimension plus importante était celle de la qualité de la langue.

2.2.2 Questionnaire de recommandations classiques

Les recommandations ont été triées par ordre d'importance, c'est-à-dire, du coefficient le plus élevé au plus faible. La moyenne globale du questionnaire pour l'échantillon des 18-35 ans a été évaluée à 5.53 (*ET* =1.77), celle pour l'échantillon 36-65 ans a évaluée à 5.46 (*ET* =1.81), et celle pour l'échantillon +65 ans a été évaluée à 5.49 (*ET* =1.8) (cf. 2.3. Grilles constituées par module et par groupe).

Le test non-paramétrique de Wilcoxon met en évidence les différences entre les dimensions ($p < .05$) pour l'échantillon global. La dimension **affichage contextuel** est jugée moins importante (*M*=4.95 ;

ET=1.13) que la dimension **qualité de la langue des sous-titres** (*M*=5.85 ; *ET*=0.867) , W=729; *p*<.001. La dimension **affichage contextuel** est jugée moins importante (*M*=4.95; *ET*=1.13) que la dimension **ergonomie des sous-titres** (*M*=5.59 ; *ET*=0.82) , W=1174; *p*<.001. La dimension **affichage contextuelle** est jugée moins importante (*M*=4.95; *ET*=1.13) que la dimension **accessibilité de l'interface** (*M*=5.67 ; *ET*=1.02) , W=1041; *p*<.001. La dimension **qualité de la langue** est jugée plus importante (*M*=5.85; *ET*=0.867) que la dimension **ergonomie du sous-titre** (*M*=5.59 ; *ET*=0.82) , W=6598; *p*<.001. La dimension **qualité de la langue** est jugée plus importante (*M*=5.59 ; *ET*=0.82) que la dimension **accessibilité de l'interface** (*M*=5.67 ; *ET*=1.02) , W=4656; *p*=0.012. Pour les dimensions **ergonomie du sous-titre** (*M*=5.59 ; *ET*=0.82) et **accessibilité de l'interface** (*M*=5.67 ; *ET*=1.02) aucune différence n'a été observée, W=3789; *p=0.104*>0.05.

Pour l'ensemble des participants, la dimension qualité de la langue était plus importante que toutes les autres dimensions. Ensuite c'était la dimension accessibilité de l'interface, suivie de l'ergonomie du sous-titre, de la qualité de la langue et enfin l'affichage contextuel.

2.2.3 Questionnaire de recommandations LSF

Les recommandations ont été triés par ordre d'importance, c'est-à -dire, du coefficient le plus élevé au plus faible. La moyenne globale du questionnaire étant évaluée à 6.56 (*ET* = 1.11). (cf. 2.3. Grilles constituées par module et par groupe).

Le test non-paramétrique de Wilcoxon met en évidence les différences entre les dimensions ($p < .05$) pour les pratiquants LSF (Tableau XXIII). La dimension **qualité de la langue** LSF est jugée plus importante (*M*=6.70 ; *ET*=0.30) que la dimension **qualité de présentation** LSF (*M*=6.52 ; *ET*=0.48), W=317 ; *p*=.032. La dimension **qualité de l'interprétation** est jugée plus importante (*M*=6.75 ; *ET*=0.154) que la dimension **qualité de présentation** LSF (*M*=6.52 ; *ET*=0.483), W=326 ; *p*=.005. La dimension **qualité de l'interprétation** est jugée plus importante (*M*=6.75 ; *ET*=0.154) que la dimension **action de l'utilisateur** (*M*=6.39 ; *ET*=0.41), W=190 ; *p*=.01. Il n'y a pas de différence entre la dimension **qualité de la langue** et la **qualité de l'interprétation** (*p*>.05).

Pour les participants, les dimensions qualité de l'interprétation et qualité de la langue LSF sont plus importantes que les dimensions qualité de présentation de la LSF et action de l'utilisateur.

2.3 Grille constituée par module et par groupe

Au total, 5 grilles d'évaluation expertes ont été constituées (1 grille technique, 1 grille utilisateurs classiques 18-35 ans, 1 grille utilisateurs classiques 36-64 ans, 1 grille utilisateurs classiques +65 ans, et 1 grille utilisateurs LSF).

2.3.1 Grille d'évaluation technique

Tableau I- Calcul des coefficients de pondération et classification des recommandations techniques

Recommandations	Coefficients de Pondération

Recommandations les plus importantes (> 5,51)	
Il faut que les téléspectateurs aient le temps nécessaire de lire les sous-titres.	7
Les sous-titres doivent au maximum faire deux lignes	6,8
Il faut respecter les règles de grammaire lors de l'écriture des sous-titres	6,64
Lorsqu'une phrase est retranscrite sur plusieurs sous-titres, son découpage doit respecter les unités de sens afin d'en faciliter sa compréhension globale	6,6
Les sous-titres doivent être organisés de manière logique sans caractères superflus	6,36
Les sous-titres doivent être positionnés dans la zone dédiée au sous-titre	6,32
Il faut essayer de réduire le plus possible le décalage entre le discours et le sous-titrage	6,08
Le nombre de caractères maximum par ligne est de quarante-deux, avec au maximum quinze caractères par seconde.	5,88
La retranscription doit veiller à ne perdre aucune information, tout en étant lisible et compréhensible	5,8
Deux sous-titres doivent être espacés d'au moins 4 images (166 ms) et terminer 2 images (80ms) avant le changement de plan	5,68
Recommandations moins importantes (< 5,51)	
L'exposition d'un sous-titre doit se situer entre 800 ms et 10 s	5,4
Chaque sous-titre doit posséder une ou plusieurs unités de sens	5,32
Les sous-titres ne doivent pas être mis durant un changement de plan	5,24
L'écart entre deux sous-titres doit se situer entre 210 et 340 ms	5,04

Pour faciliter la lecture, les sous-titres doivent être de même longueur et respecter les même règles	4,96
Les sous-titres en langue étrangère doivent être dans une autre version du film	4,52
Les polices de caractères sont calculées sur la base d'une hauteur d'image de 11 pouces et ce quel que soit le ratio.	4,44
Il est conseillé d'utiliser le format de fichier TrueType (.ttf)	2,64

2.3.2 Grille d'évaluation classique

2.3.2.1 Participants Jeunes Adultes (18 - 35 ans)

Tableau II - Calcul des coefficients de pondération et classification des recommandations pour le questionnaire classique des participants Jeunes Adultes

Recommandations	Coefficients de pondération
Recommandations les plus importantes (>5.53)	
Il faut éviter et réduire autant que possible les décalages entre le sous-titrage et les informations visuelles et sonores	6,49
Le sous-titrage doit respecter et être fidèle au sens du discours	6,42
Il faut respecter les règles d'orthographe, de grammaire et de conjugaison de la langue sous-titrée	6,24
Il faut laisser un délai d'affichage suffisant pour la lecture des sous-titres, tout en maintenant une vitesse de présentation adéquate	6,15
L'utilisateur doit pouvoir sélectionner les sous-titres et/ou la langue des signes français	6,15
Il faut respecter le vocabulaire spécifique utilisé dans la vidéo	6,07

Il faut éviter le chevauchement des sous-titres avec le texte contenu dans l'image et les autres sous-titres	6,04
Il faut activer le sous-titrage sur des lecteurs vidéo et permettre aux utilisateurs de modifier l'affichage des sous-titres.	5,98
Les instructions pour masquer ou afficher les sous-titres devraient être claires et accessibles à tout moment de la vidéo	5,98
Il faut bien isoler les sous-titres des autres contenus textuels déjà présents sur l'image	5,93
Afin de faciliter la lisibilité des sous-titres : il faut accentuer la taille des sous-titres et les contrastes (l'opposition du fond et du texte dont l'une fait ressortir l'autre)	5,91
Les sous-titres doivent toujours se situer au même endroit. De façon générale, en bas de l'image et aligné au centre. Sauf cas particuliers	5,89
Le sous-titrage doit se faire discret et respecter au mieux le rythme de montage du programme	5,60
Les sous-titres ne doivent pas utiliser de couleurs afin de ne pas être confondu avec les sous-titres sourd et malentendant	5,56
Recommandations moins importantes (<5.52)	
Les sous-titres doivent s'adapter à la taille de l'écran	5,49
Il faut respecter les codes couleurs recommandés pour le sous-titrage sourd et malentendant français (blanc = personne qui parle à l'écran, jaune = personne qui parle hors-champ, rouge= indications sonores, magenta= indications musicales, bleu cyan= voix off, vert=transcription langue étrangère)	5,42
Il faut utiliser systématiquement un tiret pour indiquer le changement de locuteur.	5,42
Le sous-titrage doit afficher les informations sonores et musicales. Par exemple, un bruit d'une porte qui claque.	5,35
Il faut distinguer les intervenants par indication des noms en début de prise de parole	5,33

Si le contraste entre le fond et l'écran n'est pas assez marqué, il faut modifier la couleur de la police (du texte) et non la place des sous-titres	5,31
Les sous-titres ne doivent pas faire plus de deux lignes	5,25
Il faut réduire la vitesse du défilement des sous-titres	5,24
Pour rendre ce qui est dit dans la vidéo encore plus clair, le texte de la vidéo pourrait être disponible pour impression ou lecture	5,09
Les sous-titres doivent être faciles à lire et à comprendre ou écrits dans un français simplifié, notamment pour les personnes avec des handicapsmentaux, les allophones, les personnes âgées et les enfants.	4,98
Il faut utiliser des majuscules lorsque le texte est dit par plusieurs personnes (un usage des majuscules pour toute autre raison est à proscrire sauf pour certains sigles et acronymes)	4,71
Afin d'améliorer les contrastes, les sous-titres doivent être écrits sur un fond gris clair	4,62
Les majuscules doivent être accentuées.	4,40
Les sous-titres doivent se positionner près de la source. Par exemple, de la personne qui parle	3,71

2.3.2.2 Participants Adultes (+35 ans et - 65 ans)

Tableau III - Calcul des coefficients de pondération et classification des recommandations pour le questionnaire classique des participants Adultes

Recommandations	Coefficients de pondération
Recommandations les plus importantes (>5.46)	
Le sous-titrage doit respecter et être fidèle au sens du discours	6,74

Il faut éviter et réduire autant que possible les décalages entre le sous-titrage et les informations visuelles et sonores	6,72
Il faut éviter le chevauchement des sous-titres avec le texte contenu dans l'image et les autres sous-titres	6,54
Il faut respecter les règles d'orthographe, de grammaire et de conjugaison de la langue sous-titrée	6,50
Il faut laisser un délai d'affichage suffisant pour la lecture des sous-titres, tout en maintenant une vitesse de présentation adéquate	6,39
Il faut bien isoler les sous-titres des autres contenus textuels déjà présents sur l'image	6,33
L'utilisateur doit pouvoir sélectionner les sous-titres et/ou la langue des signes français	6,33
Il faut activer le sous-titrage sur des lecteurs vidéo et permettre aux utilisateurs de modifier l'affichage des sous-titres.	6,15
Il faut respecter les codes couleurs recommandés pour le sous-titrage sourd et malentendant français (blanc = personne qui parle à l'écran, jaune = personne qui parle hors-champ, rouge= indications sonores, magenta= indications musicales, bleu cyan= voix off, vert=transcription langue étrangère)	6,11
Les instructions pour masquer ou afficher les sous-titres devraient être claires et accessibles à tout moment de la vidéo	6,11
Les sous-titres doivent s'adapter à la taille de l'écran	6,06
Afin de faciliter la lisibilité des sous-titres : il faut accentuer la taille des sous-titres et les contrastes (l'opposition du fond et du texte dont l'une fait ressortir l'autre)	6,00
Il faut respecter le vocabulaire spécifique utilisé dans la vidéo	5,98
Les sous-titres doivent toujours se situer au même endroit. De façon générale, en bas de l'image et aligné au centre. Sauf cas particuliers	5,91
Si le contraste entre le fond et l'écran n'est pas assez marqué, il faut modifier la couleur de la police (du texte) et non la place des sous-titres	5,76

Les sous-titres ne doivent pas faire plus de deux lignes	5,57
Le sous-titrage doit se faire discret et respecter au mieux le rythme de montage du programme	5,50
Recommandations moins importantes (<5.46)	
Il faut utiliser systématiquement un tiret pour indiquer le changement de locuteur.	5,35
Il faut distinguer les intervenants par indication des noms en début de prise de parole	5,26
Le sous-titrage doit afficher les informations sonores et musicales. Par exemple, un bruit d'une porte qui claque.	4,91
Il faut réduire la vitesse du défilement des sous-titres	4,80
Les sous-titres ne doivent pas utiliser de couleurs afin de ne pas être confondu avec les sous-titres sourd et malentendant	4,74
Afin d'améliorer les contrastes, les sous-titres doivent être écrits sur un fond gris clair	4,74
Pour rendre ce qui est dit dans la vidéo encore plus clair, le texte de la vidéo pourrait être disponible pour impression ou lecture	4,41
Les sous-titres doivent être faciles à lire et à comprendre ou écrits dans un français simplifié, notamment pour les personnes avec des handicapsmentaux, les allophones, les personnes âgées et les enfants.	4,13
Il faut utiliser des majuscules lorsque le texte est dit par plusieurs personnes (un usage des majuscules pour toute autre raison est à proscrire sauf pour certains sigles et acronymes)	4,09
Les majuscules doivent être accentuées.	3,61
Les sous-titres doivent se positionner près de la source. Par exemple, de la personne qui parle	3,24

2.3.2.3 Participants Seniors (+65 ans)

Tableau IV - Calcul des coefficients de pondération et classification des recommandations pour le questionnaire classique des participants Seniors

Recommandations	Coefficients de pondération
Recommandations les plus importantes (>5.49)	
Le sous-titrage doit respecter et être fidèle au sens du discours	6,85
Il faut éviter et réduire autant que possible les décalages entre le sous-titrage et les informations visuelles et sonores	6,79
Il faut éviter le chevauchement des sous-titres avec le texte contenu dans l'image et les autres sous-titres	6,67
Il faut bien isoler les sous-titres des autres contenus textuels déjà présents sur l'image	6,36
Il faut laisser un délai d'affichage suffisant pour la lecture des sous-titres, tout en maintenant une vitesse de présentation adéquate	6,15
Les sous-titres doivent s'adapter à la taille de l'écran	6,00
Si le contraste entre le fond et l'écran n'est pas assez marqué, il faut modifier la couleur de la police (du texte) et non la place des sous-titres	6,00
Il faut respecter les règles d'orthographe, de grammaire et de conjugaison de la langue sous-titrée	5,97
Il faut activer le sous-titrage sur des lecteurs vidéo et permettre aux utilisateurs de modifier l'affichage des sous-titres.	5,91
Les sous-titres doivent toujours se situer au même endroit. De façon générale, en bas de l'image et aligné au centre. Sauf cas particuliers	5,88
Les instructions pour masquer ou afficher les sous-titres devraient être claires et accessibles à tout moment de la vidéo	5,64

Les sous-titres ne doivent pas faire plus de deux lignes	5,64
Afin de faciliter la lisibilité des sous-titres : il faut accentuer la taille des sous-titres et les contrastes (l'opposition du fond et du texte dont l'une fait ressortir l'autre)	5,58
Il faut respecter les codes couleurs recommandés pour le sous-titrage sourd et malentendant français (blanc = personne qui parle à l'écran, jaune = personne qui parle hors-champ, rouge= indications sonores, magenta= indications musicales, bleu cyan= voix off, vert=transcription langue étrangère)	5,52
Recommandations moins importantes (<5.49)	
L'utilisateur doit pouvoir sélectionner les sous-titres et/ou la langue des signes français	5,36
Il faut respecter le vocabulaire spécifique utilisé dans la vidéo	5,33
Le sous-titrage doit se faire discret et respecter au mieux le rythme de montage du programme	5,30
Afin d'améliorer les contrastes, les sous-titres doivent être écrits sur un fond gris clair	5,24
Il faut utiliser systématiquement un tiret pour indiquer le changement de locuteur.	5,24
Les sous-titres ne doivent pas utiliser de couleurs afin de ne pas être confondu avec les sous-titres sourd et malentendant	4,85
Les sous-titres doivent être faciles à lire et à comprendre ou écrits dans un français simplifié, notamment pour les personnes avec des handicapsmentaux, les allophones, les personnes âgées et les enfants.	4,82
Il faut réduire la vitesse du défilement des sous-titres	4,79
Il faut distinguer les intervenants par indication des noms en début de prise de parole	4,73
Il faut utiliser des majuscules lorsque le texte est dit par plusieurs personnes (un usage des majuscules pour toute autre raison est à proscrire sauf pour certains sigles et acronymes)	4,73

Le sous-titrage doit afficher les informations sonores et musicales. Par exemple, un bruit d'une porte qui claque.	4,55
Pour rendre ce qui est dit dans la vidéo encore plus clair, le texte de la vidéo pourrait être disponible pour impression ou lecture	4,42
Les majuscules doivent être accentuées.	3,48
Les sous-titres doivent se positionner près de la source. Par exemple, de la personne qui parle	2,64

2.3.3 Grille d'évaluation LSF

Tableau V - Calcul des coefficients de pondération et classification des recommandations LSF

Recommandations	Coefficients de pondération
Recommandations les plus importantes (> 6,56)	
Il faut respecter le sens du message le plus fidèlement possible	6,97
Il faut avoir un interprète qualifié qui parle très bien la langue des signes	6,93
Il faut respecter les règles de l'interprétation professionnelle, par exemple rester neutre et ne pas montrer ses opinions	6,87
Il faut un bon cadrage de l'interprète idéalement cadrer jusqu'à hauteur de la mi-cuisse, avec une bonne lumière et un bon emplacement	6,87
Il faut un fond neutre et uni	6,87
Il faut diffuser l'interprétation du début à la fin	6,83
Il faut permettre à l'utilisateur de gérer l'incrustation de l'interprète sur l'écran (taille, position, etc..)	6,80

Il faut éviter la traduction "mot à mot"	6,77
Il faut incruster l'interprète idéalement sur 1/3 de l'image sur les émissions d'informations	6,77
Il faut offrir aux interprètes de bonnes conditions de travail	6,77
Il ne faut pas recouvrir le signeur avec du texte ou des informations graphiques	6,77
Il faut contraster suffisamment le fond, l'interprète et ses habits	6,73
Il faut sous-titrer les informations nécessaires à la bonne compréhension du programme lorsqu'elles ne sont pas traduites.	6,73
L'interprète doit être debout dans une bonne position	6,70
Il faut différencier les interlocuteurs en cas d'échanges complexes.	6,68
Il faut consulter les communautés concernées pour l'amélioration des services et l'évaluation des interprètes	6,67
Il faut que l'interprète respecte le secret professionnel	6,63
Il faut indiquer quand un programme est interprété en LSF et lorsqu'il ne l'est plus	6,60
Le contenu à traduire en langue des signes doit être disponible en avance pour l'interprète	6,57
Recommandations moins importantes (< 6,56)	
L'interprète doit être habillé d'une couleur neutre, unie et près du corps	6,53
Il faut respecter les règles de la langue des signes française	6,37
Il faut que les interprètes soient évaluées par des experts en langue des signes	6,20

Les contenus en LSF devraient être disponibles sur un portail spécial à la télé	5,90
Il faut incruster l'interprète dans un médaillon (un cercle) dans l'image en gros plan	5,70
Il faut avoir la LSF et les sous-titres en même temps sur le même écran	5,20

3 Synthèse évaluation technique – 2 – questionnaires en ligne – L 6.3.2

Le livrable résultat 6.3.2 a présenté les différentes analyses menées sur les prototypes ROSETTA. Ces analyses ont investigué le module sous-titrage et le module LSF. Les résultats ont concerné les réponses obtenues aux deux échelles ergonomiques élaborées dans le cadre du projet. Ces échelles ont été décrites dans le livrable méthode 6.2.4 en deux parties et proposent d'évaluer 7 composantes ergonomiques qui favorisent l'utilisabilité d'un système. Chaque module a donc été évalué selon ces composantes et les résultats sont synthétisés ci-dessous. Deux objectifs principaux sont associés au livrable 6.3.2 :

1. Positionner ROSETTA du point de vue utilisateur par rapport à l'existant.
2. Indiquer les composantes ergonomiques sur lesquelles ROSETTA peut être amélioré.

3.1 Échelle d'évaluation ergonomique en ligne

3.1.1 Description

La formulation des items de l'échelle d'évaluation étaient spécifiques à chaque module. Néanmoins ils conservaient la trame commune des critères ergonomiques à évaluer :

- L'effort cognitif (1 item)
- La compréhension (3 items)
- La comprenabilité (3 items)
- L'acceptabilité (1 item)
- L'utilisabilité (1 item)
- La satisfaction (1 item)
- L'utilité (1 item)

Soit un total de 11 items sur une échelle type Likert à 6 points qui permet aux utilisateurs d'attribuer un score de 1 à 6 pour chaque prototype présenté.

Ces deux échelles ont été extraites de questionnaires ergonomiques existants (SUS, DEEP, …) puis adaptées aux modules. L'analyse de la cohérence des échelles a montré des scores alpha cronbach>0.90. Les échelles avaient donc des qualités psychométriques très satisfaisantes (cohérence et fiabilité).

Quelques exemples d'items soumis aux utilisateurs :

1. **Effort cognitif**

a. Comprendre les signes de l'avatar m'a fatigué _code EC2

2. Compréhension

a. J'ai compris la phrase signée du premier coup _code CO1

b. J'ai dû regarder la vidéo de l'avatar à plusieurs reprises pour comprendre la phrase (-) _code RCO2

c. Je n'ai pas du tout compris la phrase (-) _ code RCO3

3.2 Synthèse Résultats phase 1 prototype sous-titres FR et LSF

3.2.1 Prototype module sous-titra FR modèle 124

3.2.1.1 Matériel

Nous avons soumis aux utilisateurs 9 vidéos à évaluer à travers l'échelle ergonomique. Les vidéos étaient composées de :

- 3 vidéos sous-titrées en FR par ROSETTA modèle 124
- 3 vidéos sous-titrées en FR par Youtube
- 3 vidéos sous-titrées en FR par un sous-titreur professionnel humain (Francetv ou freelance)

Pour chaque mode de sous-titrage, nous avons également proposé 3 catégories de vidéos :

- 1 vidéo d'Information type Journal télévisé
- 1 vidéo Divertissement type série/fiction
- 1 vidéo d'Apprentissage type Mooc/Ted Talk.

3.2.1.2 Participants

Les répondants étaient au nombre de 8 (3 femmes). La moyenne d'âge est de 41 ans (ET=14.6). 5 participants sont de langue maternelle française. 2 participants ont le Bac et les 6 autres sont de niveaux Licence – Master.

3.2.1.3 Résultats principaux

Comparaison de ROSETTA (FRV124) par rapport à l'existant

Le sous-titrage fait par un humain (traditionnel) a obtenu les meilleurs scores utilisateurs comparés aux deux autres modes de sous-titrage sur les 7 composantes ergonomiques.

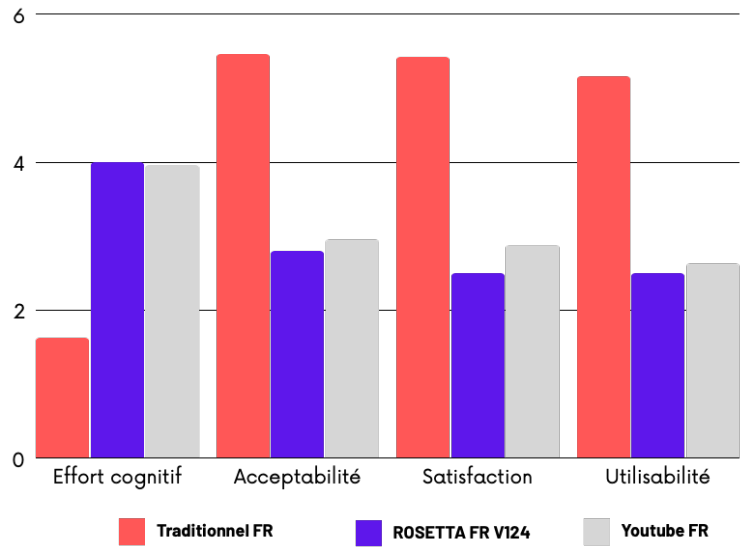

Figure 1- Moyennes des scores obtenus par les modes de sous-titrage pour les composantes ergonomiques Effort cognitif, Acceptabilité, satisfaction; Utilisabilité.

Les utilisateurs ont évalué ROSETTA (modèle FR 124) et YouTube de façon équivalente. Autrement dits, les analyses n'ont pas montré de différences au niveau des 7 composantes évaluées par les utilisateurs entre ROSETTA et Youtube. Ce qui suppose des performances ergonomiques (utilisabilité) au moins équivalentes au sous-titrage automatique disponible sur des plateformes comme Youtube.

Comparaison des catégories vidéo pour le sous-titrage ROSETTA

Les vidéos de la catégorie Apprentissage (Type Mooc/ Ted Talk) sous-titrées par ROSETTA ont obtenus de meilleurs scores que les vidéos de type Divertissement, notamment sur les composantes ergonomiques d'utilité, de satisfaction, d'utilisabilité et de comprenabilité. Ce qui suppose une meilleure appréciation par les utilisateurs de vidéos destinés à l'apprentissage, sous-titrées par ROSETTA, comparé à d'autres types de contenus.

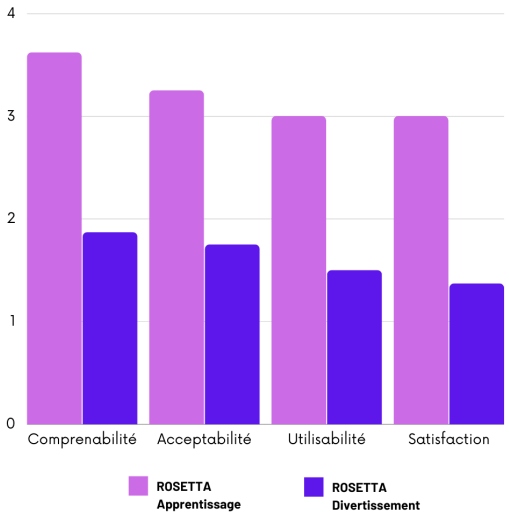

Figure 2 - Moyennes des scores obtenus par le mode de sous-titrage ROSETTA pour les composantes ergonomiques Comprenabilité, Acceptabilité, satisfaction; Utilisabilité.

3.2.2 Prototype Module LSF prototype 1

3.2.2.1 Participants

14 participants (4 hommes) ont évalué le matériel ci-dessus. La moyenne d'âge était de 36 ans (*ET*=3.5). 9 participants étaient de langue maternelle française et parmi eux 7 sourds.

3.2.2.2 Matériel

Nous avons soumis aux utilisateurs 16 phrases à évaluer à travers l'échelle ergonomique. Les phrases étaient composées de :

- 8 phrases phrases de type journalistiques générées automatiquement
    - 4 phrases générées par la méthode multicanaux, développée par ROSETTA
    - 2 phrases générées par la méthode mots isolés
    - 2 phrases générées par la méthode glose à glose
- 4 phrases issues de programmes de divertissement signées par un humain
- 4 phrases issues de programmes d'apprentissage signées par un humain

3.2.2.3 Résultats principaux

L'objectif était de tester deux hypothèses principales :

ROSETTA est-il préféré aux autres méthodes de génération?

ROSETTA fait-il mieux ou aussi bien qu'un signeur virtuel humain?

Comparaison ROSETTA aux méthodes existantes et au signeur virtuel humain

Le signeur virtuel humain était mieux compris, plus accepté, plus satisfaisant que le signeur automatique des 3 méthodes de générations. Les méthodes G2G et Mots isolés présentaient les qualités les plus pauvres.

La méthode multicanaux obtient des scores plus élevés que les deux autres méthodes de générations existantes.

Les résultats mettaient en évidence la pertinence de la méthode multicanaux privilégiée dans le projet ROSETTA pour la génération de LSF.

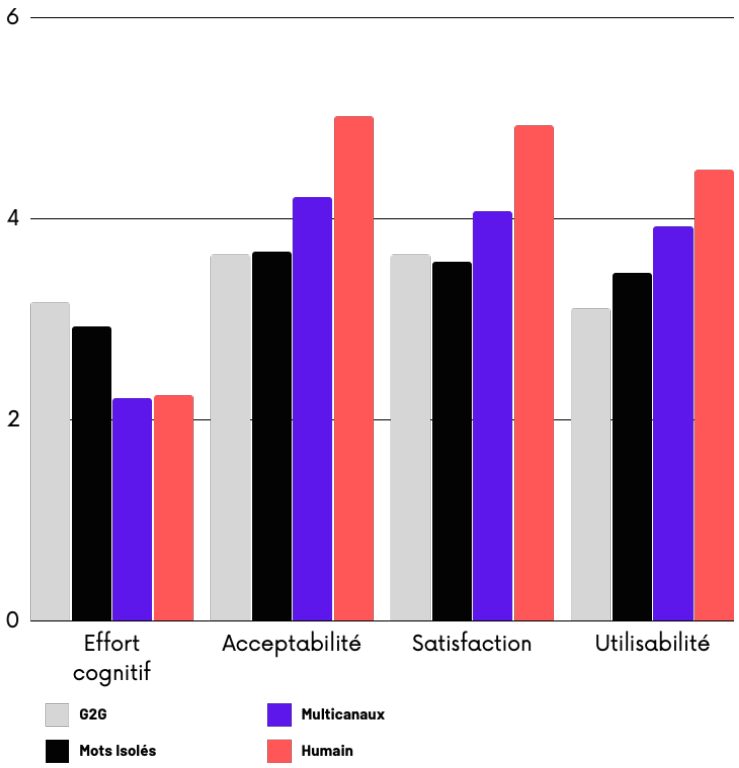

Figure 3 - Moyennes des scores obtenus par les modes de génération pour les composantes ergonomiques Effort cognitif, Acceptabilité, satisfaction; Utilisabilité.

3.2.3 Prototype module sous-titrage FR V138

3.2.3.1 Participants

Au total, 13 participants ont répondu au questionnaire (M_{age} = 34,61 ; SD = 12,90), dont 10 femmes (M_{age} = 31,6 ; SD = 11,49) et 3 hommes (M_{age} = 40,5 ; SD = 17,67). Une personne n'a pas souhaité indiquer son sexe de naissance.

3.2.3.2 Matériel

Nous avons soumis aux utilisateurs 9 vidéos à évaluer à travers l'échelle ergonomique. Les vidéos étaient composées du même matériel évalué en phase 1. Toutefois, nous avons présenté 3 vidéos sous-titrées en français par ROSETTA (modèle V138) avec un ajustement sur les vidéos présentées. La vidéo d'apprentissage a été remplacée par celle d'un TedTALK de meilleure qualité de présentation des sous-titres et une version récente d'un journal TV de France 2.

3.2.3.3 Résultats principaux

En phase 2 le sous-titrage traditionnel présentait de meilleurs résultats que les méthodes automatiques. De plus, les analyses n'ont pas permis de montrer une amélioration des performances de ROSETTA (V138) par rapport à YouTube ni par rapport à la version V124. La seule amélioration constatée entre la version 138 et la version 124 concernait la catégorie apprentissage et particulièrement la composante compréhension. Toutefois, cette amélioration pourrait être dûe au changement de matériel (vidéo mooc remplacée par une vidéo TedTAlk mieux présentée).

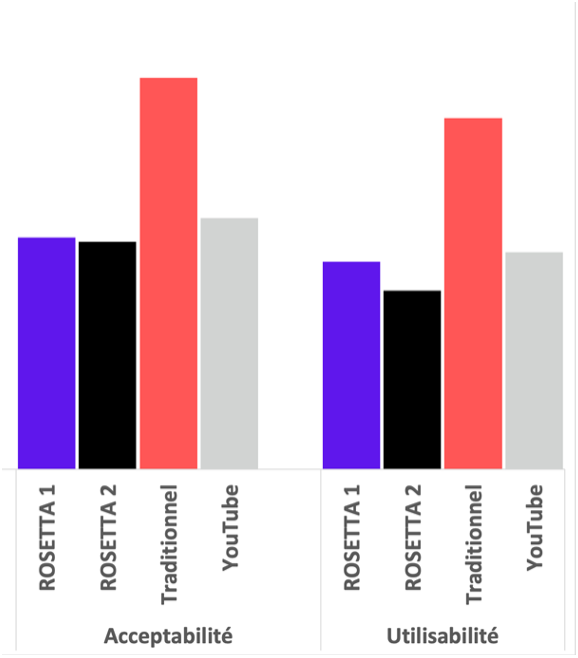

Figure 4 - Moyennes des scores obtenus par les modes de sous-titrage pour les composantes ergonomiques Effort cognitif, Acceptabilité, satisfaction; Utilisabilité.

3.2.4 Prototype module sous-titrage multilingue (Anglais) V2

3.2.4.1 Participants

Au total, 12 participants ont répondu au questionnaire (M_{age} = 37,41 ; SD = 11,64), dont 10 femmes (M_{age} = 38,6 ; SD = 12, 41) et 2 hommes (M_{age} = 31,5 ; SD = 4,94).

3.2.4.2 Matériel

Nous avons soumis aux utilisateurs 8 vidéos à évaluer à travers l'échelle ergonomique. Les vidéos étaient composées du même matériel évalué en phase 1. Toutefois, nous avons présenté 2 vidéos sous-titrées en français traditionnel, faute de journal TV français sous-titrés en anglais. La vidéo d'apprentissage a été remplacée par celle d'un TedTALK de meilleure qualité de présentation des sous-titres et une version récente d'un journal TV de France 2.

3.2.4.3 Résultats principaux

Le traditionnel présentait de meilleures qualités que les 2 modes de sous-titrages évalués. Toutefois ROSETTA (V2) présentait de meilleurs scores que le sous-titrage de YouTube, notamment sur les composantes d'effort cognitif, d'utilisabilité ou encore d'acceptabilité.

Les résultats ont également montré une relation entre le niveau de langue anglaise des participants et les scores qu'ils attribuent. Plus le niveau de langue anglaise était faible, plus les scores étaient élevés. Les francophones, apprenants anglais, ont donné une meilleure appréciation de ROSETTA comparé aux sous-titrage Youtube, notamment sur les composantes d'effort cognitif, d'acceptabilité et d'utilisabilité. Par ailleurs, les apprenants ont accordé des scores plus élevés à la vidéo d'apprentissage comparé aux vidéos d'information et de divertissement.

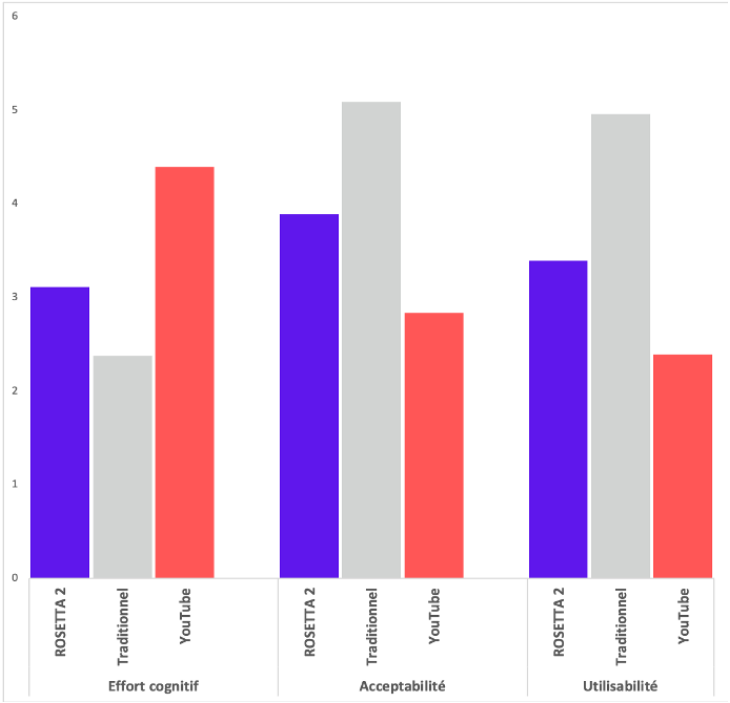

Figure 5 - Moyennes des scores obtenus par les modes de sous-titrage pour les composantes ergonomiques Effort cognitif, Acceptabilité; Utilisabilité

3.2.5 Prototype module LSF prototype 2

3.2.5.1 Participants

6 participants (5 femmes) ont répondu à la deuxième phase de tests en ligne. La moyenne d'âge est de 38 ans (*ET*=6.98). Tous les participants ont un niveau LSF avancé (C2, C1). 4 sont sourds et les 2 autres sont entendants.

3.2.5.2 Matériel

Au cours de cette deuxième phase, le matériel de génération automatique a été remplacé par du matériel multicanaux, les autres méthodes de génération moins performantes ont été exclues. Chaque phrase journalistique proposait donc un contexte image et sous-titres qui devait faciliter la compréhension.

3.2.5.3 Résultats principaux

L'objectif était de tester deux hypothèses principales :

ROSETTA prototype 1 est-il préféré à ROSETTA prototype 2 ?

ROSETTA prototype 2 fait-il mieux ou aussi bien qu'un signeur virtuel humain?

Le signeur virtuel humain présente de meilleures qualités ergonomiques comparé au signeur virtuel ROSETTA 1 et 2. En effet, le signeur virtuel humain est jugé plus comprenable, utile, utilisable et satisfaisant que le signeur virtuel en phase 2.
L'apport du contexte dans ROSETTA 2 n'a pas suffi à améliorer les performances ergonomiques attribuées en phase 1. Aucune différence n'a été constatée entre la première et la seconde version de l'avatar.

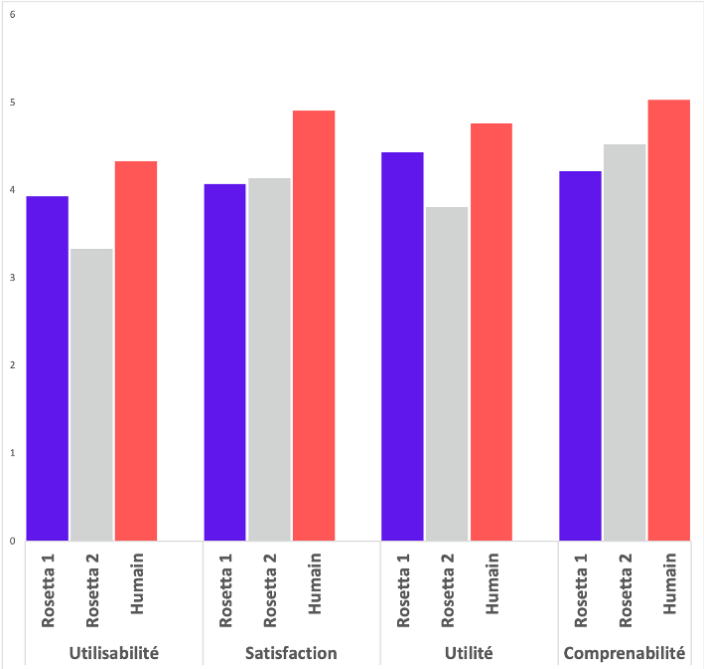

Figure 6 - Moyennes des scores obtenus par les modes de génération LSF pour les composantes ergonomiques Utilité, Comprenabilité, satisfaction; Utilisabilité.

4 Synthèse évaluation technique – 3 – Corpus bruité – L 6.3.3

4.1 Principaux résultats des analyses du corpus bruité

4.1.1 Rappel de la méthode

Les objectifs de cette étude sur le corpus bruité sont dans un premier de déterminer une méthode d'évaluation de la qualité du sous-titrage permettant la détermination de seuils d'accessibilité au contenu. Pour évaluer la qualité du sous-titrage, il est nécessaire d'avoir une métrique. Celle qui a été retenue avec le LISN est basée sur la mesure des effets du bruit introduit dans le sous-titrage. Dans un second temps, il s'agit de mettre en avant la qualité du sous-titrage en comparant sont appréciation corrélativement à celle des corpus plus ou moins bruités.

Le matériel télévisé comprend 29 vidéos différenciés selon 5 types d'émissions : une séquence, d'une durée t (environ 1mn), composant une séquence signifiante, est extraite de chaque vidéo. Cette séquence constitue un des 29 extraits qui peuvent être visualisés selon 5 niveaux de bruit correspondant au pourcentage d'erreurs (*exemple : fautes d'orthographe, lettre majuscule placée au mauvais endroit*) dans les sous-titres. Les pourcentages d'erreur utilisés étaient de 0 %, 0,5 %, 1 %, 2 % et 5 %.

Un groupe de cinq juges a été constitué pour réaliser le contrebalancement du plan en carré latin de passation. L'ordre de visualisation des extraits à présenter à un participant est aléatoire.

Le plan d'expérience est le plan P5 *E1 <T5*B5> où P5 représente les 5 juges-participants visualisant 25 extraits : un extrait (E1) de chacun des 5 types d'émission (T5) dans un des 5 niveaux de bruit (B5) avec le sous-titrage original non bruité comme ligne de base.

4.1.2 Prédictions expérimentales et Résultats attendus

L'acceptabilité de la part du téléspectateur, participant-juge, va dépendre de la compréhension qu'il a pu avoir de l'extrait vidéo plus ou moins bruité. Cette compréhension va dépendre elle-même de la lisibilité des sous-titres. Il faut donc s'attendre à ce que la lisibilité soit évaluée plus positivement que la compréhension et celle-ci plus positivement que l'acceptabilité. Ainsi, on peut trouver lisible, un sous-titre qu'on ne comprend pas et comprendre un sous-titre qu'on trouve inacceptable. Cette prédiction est retrouvée dans les évaluations des cinq juges (Figure 1 ci-dessous).

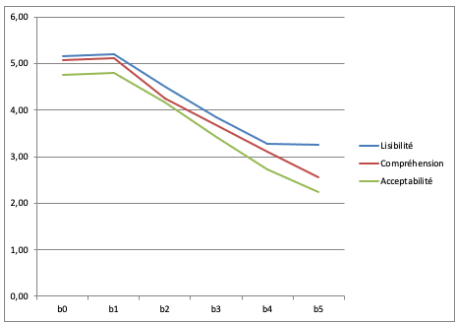

Figure 7. Effet différencié du niveau de bruit sur les trois dimensions de l'appréciation du sous-titrage. Il y a un effet du bruit (de 0 à 5 %) qui fait chuter fortement le niveau d'appréciation de l'acceptabilité, de la compréhension et de la lisibilité. Comme prédit par l'implication des dimensions, le score de lisibilité est supérieur au score de compréhension, lui-même supérieur au score d'acceptabilité. Enfin, pour ces trois dimensions, pour le sous-titrage non bruité, le score n'est pas différent de celui du sous-titrage bruité à 0,5%.

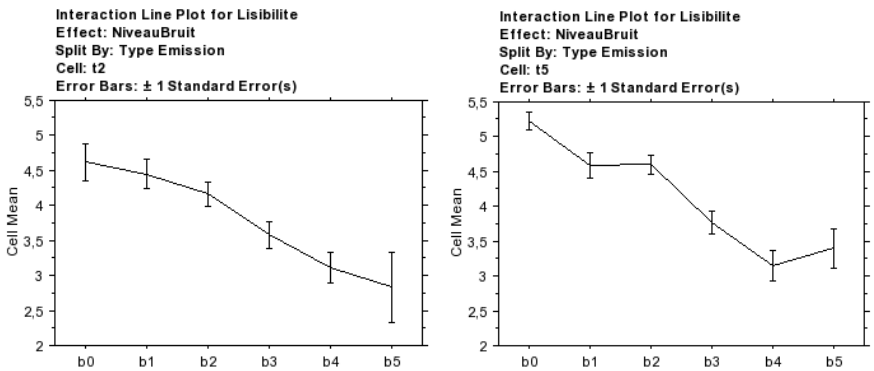

Figure 8. Effet différencié du bruit selon le type d'émission (figure de gauche : t2 - Journal télévisé vs figure de droite : t5 - Indéterminé : émissions de débats « c dans l'air ») : 'effet du bruitage dépend du type de l'émission.

4.2 Résultats produits par le bruitage sur les dimensions d'appréciation selon l'émission

L'effet du niveau de bruit dépend du type de l'émission comme montré dans la figure 2 concernant la lisibilité des sous-titrage pour le Journal télévisé (t2) comparé l'émission de débat comme « c dans l'air ». Enfin, le degré de corrélation des évaluations attribuées aux trois dimensions aumente avec le niveau de bruit (figure 3, panel du haut).

Tableau VI. Le degré de corrélation des évaluations attribuées aux trois dimensions selon le niveau de bruit (panel du haut) et selon le type de l'émission (panel du bas). Le taux de corrélation augmente avec le niveau de bruit. (figure 3, panel du haut).

Correlation Coefficient
Split By: NiveauBruit
Hypothesized Correlation = 0

	Correlation	Count	Z-Value	P-Value	95% Lower	95% Upper
Lisibilite, Comprehension: Total	,660	724	21,267	<,0001	,616	,699
Lisibilite, Comprehension: b0	,379	85	3,607	,0003	,180	,547
Lisibilite, Comprehension: b1	,521	145	6,878	<,0001	,391	,630
Lisibilite, Comprehension: b2	,530	145	7,027	<,0001	,401	,638
Lisibilite, Comprehension: b3	,594	145	8,145	<,0001	,477	,690
Lisibilite, Comprehension: b4	,564	145	7,607	<,0001	,441	,666
Lisibilite, Comprehension: b5	,575	59	4,902	<,0001	,374	,724
Lisibilite, Acceptabilite: Total	,711	724	23,900	<,0001	,673	,746
Lisibilite, Acceptabilite: b0	,562	85	5,753	<,0001	,396	,692
Lisibilite, Acceptabilite: b1	,695	145	10,220	<,0001	,600	,771
Lisibilite, Acceptabilite: b2	,528	145	7,002	<,0001	,400	,636
Lisibilite, Acceptabilite: b3	,681	145	9,906	<,0001	,583	,760
Lisibilite, Acceptabilite: b4	,601	145	8,284	<,0001	,486	,696
Lisibilite, Acceptabilite: b5	,514	59	4,249	<,0001	,297	,680
Comprehension, Acceptabilite: Total	,858	724	34,515	<,0001	,837	,876
Comprehension, Acceptabilite: b0	,833	85	10,849	<,0001	,754	,888
Comprehension, Acceptabilite: b1	,696	145	10,248	<,0001	,601	,772
Comprehension, Acceptabilite: b2	,785	145	12,598	<,0001	,713	,840
Comprehension, Acceptabilite: b3	,831	145	14,192	<,0001	,772	,875
Comprehension, Acceptabilite: b4	,821	145	13,827	<,0001	,760	,868
Comprehension, Acceptabilite: b5	,776	59	7,739	<,0001	,648	,861

Correlation Coefficient
Split By: Type Emission
Hypothesized Correlation = 0

	Correlation	Count	Z-Value	P-Value	95% Lower	95% Upper
Lisibilite, Comprehension: Total	,660	724	21,267	<,0001	,616	,699
Lisibilite, Comprehension: t1	,815	100	11,234	<,0001	,736	,872
Lisibilite, Comprehension: t2	,686	199	11,761	<,0001	,604	,753
Lisibilite, Comprehension: t3	,467	150	6,137	<,0001	,331	,584
Lisibilite, Comprehension: t4	,757	75	8,394	<,0001	,640	,840
Lisibilite, Comprehension: t5	,643	200	10,705	<,0001	,553	,717
Lisibilite, Acceptabilite: Total	,711	724	23,900	<,0001	,673	,746
Lisibilite, Acceptabilite: t1	,811	100	11,127	<,0001	,731	,869
Lisibilite, Acceptabilite: t2	,740	199	13,308	<,0001	,670	,797
Lisibilite, Acceptabilite: t3	,611	150	8,616	<,0001	,500	,703
Lisibilite, Acceptabilite: t4	,753	75	8,324	<,0001	,635	,837
Lisibilite, Acceptabilite: t5	,670	200	11,392	<,0001	,586	,740
Comprehension, Acceptabilite: Total	,858	724	34,515	<,0001	,837	,876
Comprehension, Acceptabilite: t1	,894	100	14,203	<,0001	,846	,928
Comprehension, Acceptabilite: t2	,843	199	17,225	<,0001	,797	,879
Comprehension, Acceptabilite: t3	,833	150	14,518	<,0001	,776	,876
Comprehension, Acceptabilite: t4	,877	75	11,560	<,0001	,811	,921
Comprehension, Acceptabilite: t5	,846	200	17,443	<,0001	,802	,881

5 Conclusions sur l'évaluation utilisateur des modules sous-titrage et LSF

Les modules de génération automatique de traduction en LSF et de sous-titrage français et multilingue ont été évalués tout au long du projet ROSETTA.

L'ensemble de résultats sont donnés dans les livrables suivants :

- 6.3.1 : Résultats de l'évaluation technique 1 : Recommandations pour le sous-titrage et la LSF
- 6.3.2 : Résultats de l'évaluation technique 2 : questionnaires en ligne
- 6.3.3 : Résultats de l'évaluation technique 3 : résultats sur sous-titres bruités

Ce livrable constitue une synthèse de l'ensemble des résultats du Lot 2 obtenus en utilisant les méthodes d'observation, d'expérimentation (*en présence et à distance*) et d'analyse déterminées dans le lot 6.2 et présentées dans les livrables du LOT 6.3 tout au long du projet. Ce livrable 6.3.4 est un résumé de l'ensemble des livrables du lot 6.3.

Dans ce lot 6.3., l'appréciation par les participants du module sous-titre et celle du module LSF ont été observées pour fournir des retours. Les résultats ont concerné les réponses obtenues aux deux échelles ergonomiques élaborées dans le cadre du projet. Ces échelles ont été décrites dans le livrable méthode 6.2.4 en deux parties et proposent d'évaluer 7 composantes ergonomiques qui favorisent l'utilisabilité d'un système. Chaque module a donc été évalué selon ces composantes par des utilisateurs en ligne et les résultats ont été présentés.

En phase 1, l'objectif a été de tester l'échelle de mesure élaborée pour les évaluations utilisateurs en ligne. Les analyses ont révélé une bonne fiabilité de l'échelle de mesure à 13 items. Ce nombre a été abaissé à 12 items pour augmenter d'autant plus ses qualités psychométriques.

Le second objectif a été pour le module LSF de distinguer les apports de la méthode multicanaux par rapport à d'autres méthodes de générations et également par rapport à un signeur virtuel exemplaire : l'humain.

Les vidéos du signeur virtuel humain présentent un score plus élevé dans toutes les qualités ergonomiques (effort cognitif, compréhension, comprenabilité, acceptabilité, utilisabilité, satisfaction et utilité) comparé aux méthodes de génération automatiques. De même, parmi les vidéos générées automatiquement, la méthode multicanaux présente de meilleurs scores ergonomiques. Il est intéressant de noter la nette appréciation des utilisateurs pour la méthode automatique multicanaux privilégiée dans le cadre du projet. ROSETTA (LSF) semble donc correspondre aux attentes des utilisateurs. Toutefois, les performances étant plus faibles qu'un signeur virtuel humain, ROSETTA pourra être amélioré dans le futur.

Le troisième objectif en phase a été également de tester le module sous-titrage de ROSETTA par rapport à d'autres modes de sous-titrage existants. Le modèle ROSETTA version française (124) présente des qualités ergonomiques pauvres, similaires au sous-titrage automatique de YouTube. Le sous-titrage traditionnel humain présente de biens meilleures qualités ergonomiques selon les utilisateurs.

À la suite des résultats de la phase 1, l'objectif a été de dupliquer les questionnaires en ligne sur de nouvelles versions de ROSETTA améliorées par les concepteurs. Plus spécifiquement à l'échelle, celle-ci présente de très bonnes qualités psychométriques avec un alpha de cronbach supérieur à 0.90.

Le matériel LSF évalué a subi quelques changements. Premièrement une quantité plus conséquente de vidéos pour la méthode multicanaux a été privilégiée pour les évaluations comparé aux autres méthodes de génération. Deuxièmement, les concepteurs ont voulu savoir si l'apport de contexte telle qu'une image et des sous-titres pouvaient améliorer les performances du signeur virtuel. Les résultats ont montré que le signeur virtuel humain présentait des scores plus élevés que la 2ème version multicanaux. Les analyses n'ont pas permis de mettre en évidence des différences de scores entre la version multicanaux 1 (non contextualisée) et la version 2 (contextualisée). Des analyses approfondies devraient être menées pour déterminer les axes d'amélioration prioritaires.

Par ailleurs, dans le cadre du module sous-titrage, l'intelligence artificielle de ROSETTA s'étant améliorée depuis la phase 1, il semblait important d'évaluer ses performances. Ainsi en phase 2, nous avons comparé les plus récentes versions de ROSETTA à 2 modes de sous-titrage existant. De même, nous avons comparé les nouvelles versions de ROSETTA aux plus anciennes pour mettre en évidence des signes d'amélioration. Les résultats ont montré que le traditionnel présentait des scores plus élevés que les autres modes de sous-titrage. De plus, les performances de ROSETTA français étaient supérieures à ROSETTA multilingue sauf pour la composante Utilité. Cela peut être dû à l'influence de la manipulation de deux langues qui peut accentuer l'effort cognitif quand on lit les sous-titres anglais de programmes en français. En revanche, par rapport au sous-titrage de YouTube ROSETTA Fr (138), la plus récente ne présente pas de meilleurs scores. Seule ROSETTA multilingue (anglais) présente des scores supérieurs à Youtube mais inférieurs au traditionnel. Toutefois, il faut noter le biais de niveau de compétence dans la langue anglaise semble avoir affecté les réponses. Plus le niveau des participants est faible, plus ils évaluaient les sous-titres anglais de bonne qualité.

Ces résultats posent la question des usages possibles de ROSETTA pour les apprenants d'une langue. Du fait de niveaux intermédiaires ces utilisateurs semblent moins sensibles aux erreurs de langues plus rapidement repérés par les utilisateurs dont c'est la langue maternelle. Les apprenants anglais estiment également plus utile d'avoir des programmes français sous-titrés en anglais que sous-titrés en français. De même, les résultats ont montré une pertinence de la méthode de génération LSF développée dans le projet ROSETTA en vue de projets destinés à améliorer la génération automatique LSF.

Enfin, les résultats des livrables 6.2.5 (Méthode) et 6.3.3 (Résultats) montrent que les appréciations portées selon le niveau de bruitage peuvent servir de métrique pour évaluer la qualité des sous-titrages, selon la correspondance entre appréciations comparant celles du sous-titrage cible à évaluer avec celles des corpus plus ou moins bruités.

6 Annexes

6.1 Annexe 1 : Tableau d'analyses de la cohérence des dimensions du questionnaire sous-titrage classique pour chacun des groupes.

Analyse de la cohérence des dimensions du questionnaire répartis par classes d'âges

			Cronbach's α	
	ÉCHANTILLON (N=142)	**Seniors (n=33)**	**Adultes (n=54)**	**Jeunes Adultes (n=55)**
Questionnaire classique	0.888*	0.896*	0.799*	0.924*
Dimension affichage contextuel	0.718*	0.675*	0.596	0.802*
Dimension qualité de la langue	0.353	0.480	0.334	0.330
Dimension ergonomie du sous-titre	0.781*	0.769*	0.617	0.855*
Dimension accessibilité de l'interface	0.565	0.602	0.334	0.674*