Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Mesures linguistiques automatiques pour l'évaluation des systèmes de Reconnaissance Automatique de la Parole

Thibault Roux[1, 2] Mickaël Rouvier[2] Jane Wottawa[3] Richard Dufour[1]
(1) LS2N, 2 chemin de la Houssinière, 44300 Nantes, France
(2) LIA, 339 chemin des Meinajaries, 84000 Avignon, France
(3) LIUM, avenue Olivier Messiaen, 72085 Le Mans, France
thibault.roux@univ-nantes.fr, richard.dufour@univ-nantes.fr,
mickael.rouvier@univ-avignon.fr, jane.wottawa@univ-lemans.fr

RÉSUMÉ

L'évaluation de transcriptions issues de systèmes de Reconnaissance Automatique de la Parole (RAP) est un problème difficile et toujours ouvert, qui se résume généralement à ne considérer que le WER. Nous présentons dans cet article un ensemble de métriques, souvent utilisées dans d'autres tâches en traitement du langage naturel, que nous proposons d'appliquer en complément du WER en RAP. Nous introduisons en particulier deux mesures considérant les aspects morpho-syntaxiques et sémantiques des mots transcrits : 1) le POSER (Part-of-speech Error Rate), qui évalue les aspects grammaticaux, et 2) le EmbER (Embedding Error Rate), une mesure originale qui reprend celle du WER en apportant une pondération en fonction de la distance sémantique des mots mal transcrits. Afin de montrer les informations supplémentaires qu'elles apportent, nous proposons également une analyse qualitative décrivant l'apport au niveau linguistique de modèles de langage utilisés pour le réordonnancement d'hypothèses de transcription a posteriori.

ABSTRACT

Automated linguistic measures for automatic speech recognition systems' evaluation

Evaluating transcriptions from automatic speech recognition (ASR) systems is a difficult and still open problem, which often boils down to not considering only the word-error rate (WER). We present in this article a set of metrics, often used in other tasks in natural language processing (NLP), which we propose to apply in addition to WER in ASR. In particular, we introduce two measures relating to the morpho-syntactic and semantic aspects of transcribed words : 1) the POSER (Part-of-speech Error Rate), which highlights the grammatical aspects, and 2) the EmbER (Embedding Error Rate), an original measurement which takes up that of the WER by providing a weighting according to the semantic distance of the badly transcribed words. In order to show the additional information they provide, we also offer a qualitative analysis describing the contribution at the linguistic level of the language models used for the a posteriori rescoring of transcription hypotheses.

MOTS-CLÉS : Reconnaissance automatique de la parole, métriques d'évaluation, plongements lexicaux, distance sémantique, étiquetage morpho-syntaxique.

KEYWORDS: Automatic speech recognition, evaluation metrics, embeddings, semantic distance, part-of-speech tagging.

1 Introduction

Grâce aux avancées scientifiques et technologiques de ces dernières années, les systèmes de Reconnaissance Automatique de la Parole (RAP), qui permettent de transcrire de la parole en texte, sont maintenant utilisés dans de nombreuses applications liées au Traitement Automatique du Langage

Volume 1 : conférence principale, pages 166–173.

(TAL). Les systèmes de RAP ont notamment profité de l'augmentation massive des données disponibles, en particulier pour l'apprentissage de leurs modèles (Baevski *et al.*, 2020), ainsi que des approches par apprentissage profond (Deng *et al.*, 2013; Amodei *et al.*, 2016).

D'un point de vue applicatif, plusieurs contextes d'utilisation sont possibles : une transcription automatique peut être soit utilisée directement (*e.g.* pour le sous-titrage automatique), soit être une partie (souvent en entrée) d'un autre système (*e.g.* dialogue humain-machine, indexation automatique de documents audio, etc.). Malgré les avancées actuelles, des erreurs dans les transcriptions automatiques sont inévitables et impactent son utilisation : par exemple, les erreurs de transcription peuvent se répercuter sur les applications qui les utilisent, et donc influer négativement sur leurs performances, ou encore rendre la compréhension des transcriptions difficile par un humain.

Les systèmes de RAP sont principalement évalués avec la métrique du taux d'erreur-mot (WER pour *Word Error Rate*). Cette métrique possède l'avantage d'être simple à mettre en place, car elle ne nécessite d'avoir qu'une transcription de référence (*i.e.* annotée manuellement) des mots. Elle est néanmoins limitée dans le sens où aucune autre information que le mot lui-même n'est intégrée (*e.g.* aucune information linguistique n'est prise en compte, pas de connaissance sémantique, etc.). Chaque erreur a également le même poids au sein de cette métrique alors même que l'on sait, sachant la tâche visée, que les mots ont une importance différente au sein d'une transcription textuelle (Morchid *et al.*, 2016). Ces limites ont déjà été exposées dans le passé, avec des variantes proposées comme par exemple le IWER (Mdhaffar *et al.*, 2019), qui se concentre notamment sur les mots choisis comme *importants* dans une transcription.

Dans cet article, nous étudions un ensemble de mesures automatiques pour aider à l'évaluation des systèmes de RAP, en particulier sur les aspects liés au langage. Ces mesures doivent permettre une analyse plus fine des erreurs de transcription, en mettant en avant certains aspects des erreurs (classes morphosyntaxiques, erreurs en contexte, distance sémantique, etc.). Un des avantages de ces mesures est qu'elles ne nécessitent aucune annotation manuelle supplémentaire des transcriptions et peuvent s'appliquer à n'importe quelle langue. De plus, leur multiplication permet de mettre en avant des visions différentes des erreurs, ces métriques pouvant alors se compléter. Nous proposons ensuite une analyse qualitative au moyen de ces mesures sur un système de RAP à l'état de l'art, en décrivant plus finement l'apport du réordonnancement des hypothèses a posteriori (*rescoring*) par un modèle de langage (ML) quadri-gramme couplé à un ML utilisant des réseaux de neurones récurrents (RNNLM).

L'article est organisé comme suit. Dans la partie 2, nous décrivons la métrique classique du WER, puis nous listons et détaillons les différentes mesures automatiques que nous proposons pour permettre une évaluation plus fine des transcriptions au niveau linguistique. Afin de comprendre l'intérêt de l'utilisation de ces mesures, une analyse qualitative est proposée, en détaillant tout d'abord le protocole expérimental dans la partie 3, puis les résultats et analyses dans la partie 4. Enfin, une conclusion ainsi que des perspectives sont fournies dans la partie 5.

2 Description des mesures automatiques proposées

Les systèmes de RAP sont principalement évalués au travers du WER. Dans cette partie, nous proposons tout d'abord de la décrire (sous-partie 2.1) afin de prendre conscience de ses avantages et limites. Puis nous détaillons les cinq mesures automatiques complémentaires que nous souhaitons appliquer à l'évaluation des transcriptions automatiques au niveau linguistique (sous-parties 2.2, 2.3, 2.4, 2.5 et 2.6) en complément du WER.

2.1 Taux d'erreur-mot (WER)

Il s'agit ici de comparer une transcription de référence (manuelle) avec une transcription automatique obtenue avec un système de RAP. La métrique du WER prend alors simplement en compte trois types d'erreurs :

— *Substitution (Sub)* : mot reconnu en lieu et place de la transcription de référence.
— *Insertion (Ins)* : mot transcrit inséré par rapport à la transcription de référence.
— *Suppression (Del)* : mot de la référence non transcrit par le système de RAP.

Les phrases suivantes en exemple proposent un alignement entre une phrase de référence (*Référence*) et une transcription automatique (*Hypothèse*) permettant le calcul du WER :

Référence	Tu	ne	manges		pas	ton	kiwi
	=	=	*Sub*	*Ins*	=	*Sub*	*Del*
Hypothèse	Tu	ne	mens	je	pas	toi	

Formellement, le WER se calcule alors comme suit :

$$WER = \frac{\#Sub + \#Ins + \#Del}{\#mots\ de\ la\ référence} \tag{1}$$

Par définition, le WER considère donc tout type d'erreur d'une importance équivalente. Cela constitue l'avantage principal de cette métrique : sa simplicité d'application et d'utilisation. Mais le WER souffre néanmoins de limites. En reprenant l'exemple précédent, le mot *manges* a été transcrit par le mot *mens*. Une autre hypothèse de transcription aurait pu être *mange*, qui est ici mal accordé. Dans les deux cas, le WER serait identique par rapport à la référence, alors même que la nature de l'erreur est différente. Une autre limitation concerne le peu de catégories prises en compte pour le calcul du taux. De même, ces catégories se limitent aux mots transcrits, sans aucune information supplémentaire.

2.2 Taux d'erreur-caractère (CER)

Le taux d'erreur caractère (CER pour *Character Error Rate*) s'appuie sur le même principe que le WER mais appliqué aux caractères. Il a déjà été utilisé dans le domaine de la RAP (Xu *et al.*, 2021). Initialement, il convient particulièrement aux langues fondées sur les caractères comme les langues asiatiques. Pour les langues latines, et en particulier le français, le CER permet, entre autres, de donner une indication quant à la nature des erreurs : un CER bas pourrait indiquer que le système de RAP a tendance à générer des mots proches du mot de référence (et donc potentiellement intégrer des erreurs liées au genre, au nombre, au temps, etc.) contrairement à un CER élevé, avec des hypothèses de transcription très éloignées.

2.3 Taux d'erreur des classes morpho-syntaxiques (POSER)

Nous avons également choisi d'utiliser une métrique permettant le calcul du taux d'erreur sur les classes morpho-syntaxiques d'une transcription (POSER pour *Part-of-speech Error Rate*). Le POSER nous permet de savoir si les phrases transcrites sont grammaticalement proches de celles de référence, et de mieux caractériser les erreurs de substitution. Ce taux est calculé avec la même formule que le WER, sauf que ce sont les classes morpho-syntaxiques qui sont prises en compte au lieu des mots.

2.4 Taux d'erreur par plongements lexicaux (EmbER)

Le sens des mots n'est pas considéré dans la métrique du WER. Pour remédier à cela, nous proposons d'utiliser une métrique fondée sur les plongements lexicaux des mots (*word embeddings*). De la même manière que Le *et al.*, nous avons pour objectif de garder la métrique du WER mais en la pondérant : un mot n'est plus considéré de façon binaire (0 pour bien transcrit et 1 pour une erreur), les erreurs étant pondérées selon leur distance sémantique par rapport au mot de référence. Cette

distance est calculée à l'aide de la similarité cosinus entre le plongement du mot de la référence et celui du mot transcrit substitué. Dans le cadre de nos expériences, nous avons utilisé les plongements lexicaux de Fasttext (Bojanowski *et al.*, 2017) et appliquons une erreur de 0,1 si la similarité cosinus est au-dessus d'un seuil de 0,4, et de 1 dans les autres cas.

2.5 Distance sémantique de mots (BERTScore)

Développé pour la génération de texte (Zhang *et al.*, 2019), cette métrique a pour objectif de comparer un mot référence et une hypothèse par rapport à une proximité sémantique. La première étape consiste à obtenir les mots et sous-mots (*tokens*) de la référence et de l'hypothèse grâce au tokenizer WordPiece utilisé par BERT (Devlin *et al.*, 2018). Ensuite, étant donné les séquence de plongements contextualisés de référence ($x1$,...,xk) et d'hypothèse ($\hat{x}1$,...,$\hat{x}m$), la similarité cosinus est calculée entre chaque plongement de la référence et de l'hypothèse pour obtenir une matrice de score pondérée ici avec la fréquence inverse du document (Zhang *et al.*, 2019).

Pour calculer la précision, on associe chaque token x à un token $\hat{x}$ en sélectionnant le token apportant la plus haute similarité. Puis le rappel est calculé en associant chaque token $\hat{x}$ à un token x de la même façon. Ensuite, le score de f-mesure, que nous utilisons dans nos expériences, est calculé à partir du rappel et de la précision (Zhang *et al.*, 2019).

2.6 Distance sémantique de phrases (SemDist)

Alors que les métriques précédentes se concentrent sur les mots et caractères, le principe de cette mesure (Kim *et al.*, 2021) est de prendre en considération la phrase complète. Dans le cadre de la RAP, la référence et l'hypothèse sont respectivement transformées en leurs plongements de phrases à l'aide d'un modèle de type SentenceBERT (Reimers & Gurevych, 2019), c'est-à-dire un modèle de plongements de phrases utilisant les plongements contextuels de mots de BERT (Devlin *et al.*, 2018). Il est alors possible de comparer ces vecteurs avec la similarité cosinus. Notre mesure finale est la moyenne des similarités cosinus entre le plongement de phrase de chaque référence et de son hypothèse respective.

3 Protocole expérimental

Nous présentons dans cette partie le protocole expérimental mis en place pour appliquer les différentes métriques listées dans la partie 2. Nous décrivons les données utilisées pour notre analyse qualitative de l'apport du *rescoring* de modèles de langage sur le français dans la partie 3.1, puis le système de RAP ainsi que l'étiqueteur morpho-syntaxique dans les parties 3.2 et 3.3 respectivement.

3.1 Données

Les jeux de données utilisés pour apprendre le système de RAP sont les corpus d'apprentissage ESTER 1 et 2 (Galliano *et al.*, 2006, 2009), EPAC (Esteve *et al.*, 2010), ETAPE (Gravier *et al.*, 2012), REPERE (Giraudel *et al.*, 2012) et des données internes au LIA. L'ensemble des corpus représentent environ 940 heures de données audios et sont des données de diffusion radiophonique et télévisée. L'évaluation des systèmes est faite sur le corpus de test REPERE, soit environ 10h de données audios.

3.2 Système de reconnaissance automatique de la parole

Le système de RAP s'appuie sur une recette existante à l'état de l'art [1] qui utilise la boîte à outils Kaldi (Povey *et al.*, 2011).

1. https://github.com/kaldi-asr/kaldi/blob/master/egs/librispeech/s5/

Le modèle acoustique est un réseau de neurones profond fondé sur l'architecture TDNN-F (Povey *et al.*, 2018). Afin de rendre le système plus robuste aux différentes conditions acoustiques, les fichiers audios ont été aléatoirement perturbés en vitesse et en volume pendant le processus d'entraînement.

Trois modèles de langage sont utilisés. Le premier est un modèle tri-gramme entraîné avec SRILM (Stolcke, 2002) et utilisé directement par le système de RAP. Le deuxième est un réseau de neurones profond de type RNNLM, prévu pour le réordonnancement des hypothèses a posteriori (*rescoring*), donc non inclus dans le système de RAP initial. Le réseau est composé de trois couches TDNN entrecoupées de deux couches LSTM. De même, un modèle quadri-gramme est utilisé pendant cette étape de *rescoring*. Cette étape optionnelle permet de recalculer les hypothèses proposées par le système de RAP de base afin d'améliorer les performances. Le corpus d'apprentissage ainsi que le vocabulaire utilisés pour apprendre le modèle tri-gramme, et les modèles RNNLM et quadri-gramme en *rescoring*, sont identiques.

3.3 Étiqueteur morpho-syntaxique

Les classes morpho-syntaxiques (POS) pour le français ont été obtenues automatiquement avec l'étiqueteur POET [2], s'appuyant sur les plongements contextuels Flair (Akbik *et al.*, 2018). Nous avons choisi cet étiqueteur car il permet d'avoir à la fois les classes génériques d'Universal Dependency (nom, adjectif, adverbe, etc.) mais également une granularité fine grâce à des informations complémentaires sur ces mêmes étiquettes (nom féminin pluriel, pronom personnel 3ème personne du singulier, etc.). Nous proposons alors deux mesures fondées sur les POS tags : une intégrant les classes détaillées (dPOSER) et une avec les classes génériques d'Universal Dependency (uPOSER). Notons qu'aucune annotation manuelle en POS tag n'a été utilisée : les transcriptions de référence comme celles d'hypothèse ont été automatiquement étiquetées.

4 Analyse des résultats

Nous proposons une étude qualitative de l'impact, au niveau linguistique, du réordonnancement des hypothèses a posteriori (*rescoring*) d'un système de RAP. Pour ce faire, nous utilisons les 5 métriques proposées dans la partie 2 en complément du WER. Nous étudions tout d'abord l'impact de ce *rescoring* dans la sous-partie 4.1 puis proposons une analyse de ces métriques dans la sous-partie 4.2.

4.1 Impact du réordonnancement des hypothèses

Le Tableau 1 présente les résultats avec les différentes métriques étudiées sur les transcriptions automatiques obtenues au moyen du système de RAP sans (Base) et avec réordonnancement des hypothèses (Rescoring). Comme attendu, ce *rescoring* permet d'améliorer les résultats quelle que soit la métrique considérée (ici, baisse des taux d'erreur) : une amélioration est donc visible au niveau des mots, des caractères, de la grammaire et de la sémantique. Les gains pour chaque métrique sont également fournis dans le Tableau 1. Ils mettent surtout en avant que les gains en relatif obtenus sur le WER sont inférieurs sur les autres métriques. Ainsi, en proportion, les métriques SemDist et BERTScore ont les gains relatifs les plus faibles, ce qui tend à nous faire dire que le *rescoring* ne corrige qu'en partie les mots transcrits qui étaient *sémantiquement* éloignés de leur référence.
La baisse observée avec la métrique EmbER, qui s'intéresse au sens véhiculé dans les phrases, nous a ensuite poussé à étudier plus en détails les segments transcrits. Ainsi, nous constatons que pour 50 % des phrases, peu importe la métrique, le modèle avec *rescoring* obtient de meilleures performances que sans (Base). Au contraire, pour 37 % des phrases, le *rescoring* obtient des performances plus faibles (aucun changement pour les 13 % restantes). Ces nouvelles informations nuancent l'amélioration de nos systèmes, car malgré une amélioration globale (Tableau 1), le *rescoring* implique aussi une baisse

2. https://huggingface.co/qanastek/pos-french

Système	WER	CER	dPOSER	uPOSER	SemDist	BERTScore	EmbER
Base	15,45	8,57	14,59	12,22	7,89	9,12	12,33
Rescoring	13,24	7,70	12,51	10,79	7,18	8,38	10,79
Réduction	-14,3 %	-10,2 %	-14,3 %	-11,7 %	-9,0 %	-8,1 %	-12,5 %

TABLE 1 – Comparaison des performances des systèmes de RAP de base (Base) et avec réordonnancement des hypothèses (Rescoring) au moyen de différentes métriques. La réduction observée entre les deux systèmes, en valeur relative, est également fournie.

de la qualité des transcriptions pour certaines phrases. Il serait donc intéressant de pouvoir comparer les gains selon les métriques, en particulier pour les parties où le WER est dégradé. Il est possible que l'amélioration ou la dégradation locale des performances soient dues à divers facteurs tels que le domaine de la parole (s'il est proche de l'écrit ou de la parole spontanée), la qualité du son, le bruit, la présence d'acronymes, de mots étrangers, etc...

4.2 Analyse des métriques

En réutilisant les mesures locales précédentes et dans le but de faire une analyse plus approfondie de nos métriques, nous avons calculé une corrélation de Pearson entre nos métriques pour nos deux systèmes. Nous avons ensuite calculé la moyenne des corrélations des deux systèmes dans le Tableau 2.

	WER	CER	dPOSER	uPOSER	SemDist	EmbER	BERTScore
WER	-	90,37	92,96	90,40	71,81	96,51	74,63
CER	90,37	-	90,61	91,29	66,79	92,05	75,45
dPOSER	92,96	90,61	-	97,95	65,33	91,00	74,09
uPOSER	90,40	91,29	97,95	-	64,13	88,98	74,25
SemDist	71,81	66,79	65,33	64,13	-	75,73	63,35
EmbER	96,51	92,05	91,00	88,98	75,73	-	84,51
BERTScore	74,63	75,45	74,09	74,25	63,35	84,51	-

TABLE 2 – Moyennes des corrélations de Pearson inter-métriques. Pour plus de lisibilité, les valeurs sont multipliées par 100.

Les métriques ne corrèlent pas toutes entre elles de la même façon. SemDist est la métrique qui corrèle le moins avec les autres. Cela s'explique par le fait qu'elle est la seule métrique fondée sur les plongements de phrase dans nos expériences, dépassant la dimension *mot*. Cette faible corrélation implique que pour des tâches en aval utilisant des plongements de phrase, l'impact sera plus faible que celui observé sur le WER. Cette idée va dans le sens de nombreuses publications dans le TAL et la RAP qui considèrent que les évaluations intrinsèques sont moins pertinentes que les évaluations extrinsèques (Wang *et al.*, 2003; Glavaš *et al.*, 2019). D'ailleurs, les auteurs de SemDist (Kim *et al.*, 2021) concluaient que leur métrique corrélait mieux avec les tâches en aval que le WER.

On peut constater que la métrique qui corrèle le mieux avec BERTScore et SemDist est l'EmbER, toutes trois fondées sur les plongements, alors que la métrique qui corrèle le mieux avec l'EmbER est le WER. Cela met en évidence que l'Embedding Error Rate est une métrique hybride qui a l'avantage de corréler au WER et aux métriques s'appuyant sur les plongements lexicaux.

5 Conclusions et Perspectives

Dans cette étude, nous avons appliqué différentes mesures en complément de la métrique WER aux systèmes de RAP afin de faire apparaître différentes dimensions linguistiques (grammaticale, sémantique, etc.) aux erreurs de transcription. En particulier, nous avons proposé une nouvelle mé-

trique, appelée EmbER, qui calcule un WER pondéré selon la similarité sémantique sur les erreurs de substitution. Nous avons choisi de vérifier leur pertinence en étudiant l'impact du réordonnancement des hypothèses a posteriori sur les systèmes de RAP au moyen de modèles de langage. Notre étude montre que les gains ne sont pas équivalents selon la métrique considérée, mettant alors en lumière les limites du WER seul pour étudier les améliorations au niveau lexical, grammatical ou encore sémantique. Dans la continuité de ces travaux, nous souhaitons étendre cette analyse en combinant les mesures. En effet, nous nous sommes ici intéressés à ces métriques de manière isolée, mais il semble pertinent d'étudier, par exemple, les mesures sémantiques sur des classes morpho-syntaxiques identifiées (par exemple, comparer BERTScore sur les noms de personnes et les adjectifs). Il serait intéressant d'intégrer les différentes métriques utilisées comme une fonction de perte pour évaluer les éventuelles améliorations des performances. À plus long terme, il serait intéressant d'évaluer la corrélation entre nos métriques et la perception humaine des erreurs.

Références

AKBIK A., BLYTHE D. & VOLLGRAF R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, p. 1638–1649.

AMODEI D., ANANTHANARAYANAN S., ANUBHAI R., BAI J., BATTENBERG E., CASE C., CASPER J., CATANZARO B., CHENG Q., CHEN G. *et al.* (2016). Deep speech 2 : End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, p. 173–182 : PMLR.

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, **33**, 12449–12460.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

BOJANOWSKI P., GRAVE É., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146.

DENG L., HINTON G. & KINGSBURY B. (2013). New types of deep neural network learning for speech recognition and related applications : An overview. In *2013 IEEE international conference on acoustics, speech and signal processing*, p. 8599–8603 : IEEE.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

ESTEVE Y., BAZILLON T., ANTOINE J.-Y., BÉCHET F. & FARINAS J. (2010). The epac corpus : manual and automatic annotations of conversational speech in french broadcast news. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

GALLIANO S., GEOFFROIS E., GRAVIER G., BONASTRE J.-F., MOSTEFA D. & CHOUKRI K. (2006). Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *LREC*, p. 139–142.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*.

GIRAUDEL A., CARRÉ M., MAPELLI V., KAHN J., GALIBERT O. & QUINTARD L. (2012). The repere corpus : a multimodal corpus for person recognition. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 1102–1107.

GLAVAŠ G., LITSCHKO R., RUDER S. & VULIĆ I. (2019). How to (properly) evaluate cross-lingual word embeddings : On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 710–721.

GRAVIER G., ADDA G., PAULSSON N., CARRÉ M., GIRAUDEL A. & GALIBERT O. (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 114–118.

KIM S., LE D., ZHENG W., SINGH T., ARORA A., ZHAI X., FUEGEN C., KALINLI O. & SELTZER M. L. (2021). Evaluating user perception of speech recognition system quality with semantic distance metric. *arXiv preprint arXiv :2110.05376*.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LE N.-T., SERVAN C., LECOUTEUX B. & BESACIER L. (2016). Better evaluation of asr in speech translation context using word embeddings. In *Interspeech 2016*.

MDHAFFAR S., ESTÈVE Y., HERNANDEZ N., LAURENT A., DUFOUR R. & QUINIOU S. (2019). Qualitative evaluation of asr adaptation in a lecture context : Application to the pastel corpus. In *INTERSPEECH*, p. 569–573.

MORCHID M., DUFOUR R. & LINARÈS G. (2016). Impact of word error rate on theme identification task of highly imperfect human–human conversations. *Computer Speech & Language*, **38**, 68–85.

POVEY D., CHENG G., WANG Y., LI K., XU H., YARMOHAMMADI M. & KHUDANPUR S. (2018). Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Interspeech*, p. 3743–3747.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P. *et al.* (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, volume CONF : IEEE Signal Processing Society.

REIMERS N. & GUREVYCH I. (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv :1908.10084*.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

STOLCKE A. (2002). Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

WANG Y.-Y., ACERO A. & CHELBA C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*, p. 577–582 : IEEE.

XU M., LI S. & ZHANG X.-L. (2021). Transformer-based end-to-end speech recognition with local dense synthesizer attention. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5899–5903 : IEEE.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2019). Bertscore : Evaluating text generation with bert. *arXiv preprint arXiv :1904.09675*.