# Qualitative Evaluation of Language Model Rescoring in Automatic Speech Recognition

Thibault Bañeras Roux, Mickael Rouvier, Jane Wottawa, Richard Dufour

▶ **To cite this version:**

Thibault Bañeras Roux, Mickael Rouvier, Jane Wottawa, Richard Dufour. Qualitative Evaluation of Language Model Rescoring in Automatic Speech Recognition. Interspeech, Sep 2022, Incheon, South Korea. hal-03712735v2

**HAL Id: hal-03712735**
**https://hal.science/hal-03712735v2**

Submitted on 17 Nov 2022

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

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

# Qualitative Evaluation of Language Model Rescoring in Automatic Speech Recognition

*Thibault Bañeras-Roux*[1], *Mickaël Rouvier*[2], *Jane Wottawa*[3], *Richard Dufour*[1]

[1]LS2N - Nantes University (France)
[2]LIA - Avignon University (France)
[3]LIUM - Le Mans University (France)

thibault.roux@univ-nantes.fr, mickael.rouvier@univ-avignon.fr,
jane.wottawa@univ-lemans.fr, richard.dufour@univ-nantes.fr

## Abstract

Evaluating automatic speech recognition (ASR) systems is a classical but difficult and still open problem, which often boils down to focusing only on the word error rate (WER). However, this metric suffers from many limitations and does not allow an in-depth analysis of automatic transcription errors. In this paper, we propose to study and understand the impact of rescoring using language models in ASR systems by means of several metrics often used in other natural language processing (NLP) tasks in addition to the WER. In particular, we introduce two measures related to morpho-syntactic and semantic aspects of transcribed words: 1) the POSER (Part-of-speech Error Rate), which should highlight the grammatical aspects, and 2) the EmbER (Embedding Error Rate), a measurement that modifies the WER by providing a weighting according to the semantic distance of the wrongly transcribed words. These metrics illustrate the linguistic contributions of the language models that are applied during a posterior rescoring step on transcription hypotheses.

**Index Terms**: Automatic speech recognition, Semantic analysis, Language modeling, evaluation metrics

## 1. Introduction

Over the last years, various speech and language processing fields have made significant progress thanks to scientific and technological advances. Automatic Speech Recognition (ASR) has notably benefited from the massive increase in available data and the use of deep learning approaches [1, 2], making its models more robust and efficient [3]. From an application point of view, several usage contexts are possible: an automatic transcription can either be used directly (*e.g.* for automatic subtitling), or it can be part (often as an input) of another application (*e.g.* human-computer dialogue, automatic indexing of audio documents, etc.). Despite the current performance, errors in automatic transcriptions are inevitable and impact its use: for example, ASR errors can affect applications where these systems are implemented, and thus negatively influence their global performance by making it difficult for humans to understand the transcriptions.

ASR systems are widely evaluated with the Word Error Rate (WER) metric. The simplicity of this metric is its main advantage and the reason of its massive adoption, as it only requires a reference transcription (*i.e.* manually annotated) of the words. It is nevertheless limited in the sense that no other information than the word itself is integrated (*e.g.* no linguistic information is taken into account, no semantic knowledge, etc.). Each error also has the same weight within this metric even though we know that words have a different impact considering a targeted task [4]. These limitations have already been exposed in the past, with proposed variants such as the IWER [5], which focuses on words chosen as *important* within a transcription.

In this paper, we investigate a set of automatic measures used in various natural language processing (NLP) tasks to help in the specific evaluation of ASR systems, especially on language-related aspects. These measures should allow for a finer analysis of transcription errors, by highlighting certain forms of the errors (part-of-speech, context errors, semantic distance, etc.). One of the advantages of these proposed measures is that they do not require any additional manual annotation of transcriptions and can be applied to any language. Moreover, their multiplication allows us to put forward different visions of the errors, these metrics can then complement each other. We then propose a qualitative analysis using these metrics on a state-of-the-art ASR system, by analyzing in more details the contribution of a posteriori reordering of transcription hypothesis, a process called rescoring, performed with a quadrigram language model (LM) coupled to a Recurrent Neural Net Language Model (RNNLM) on a French dataset.

This paper is organized as follows: in Section 2, we describe the classical WER metric, before listing and detailing the different automatic measures we propose to allow a finer evaluation of transcriptions at a linguistic level. In order to understand the interest of these measures, a qualitative analysis of language model rescoring is proposed, first detailing the experimental protocol in Section 3, then the results and analysis in Section 4. Finally, a conclusion as well as perspectives are provided in Section 5.

## 2. Description of proposed measures

ASR systems are mainly evaluated through the WER. In this section, we first describe it (Section 2.1) in order to highlight its advantages and limitations. Then we detail the 6 complementary automatic measures that we wish to apply to the evaluation of automatic transcriptions at the syntactic (Sections 2.2, 2.3 and 2.4) and semantic (Sections 2.5, 2.6 and 2.7) levels in addition to the WER.

### 2.1. Word Error Rate (WER)

This metric compares a reference (manual) transcription with an automatic transcription obtained with an ASR system on the word level, words being a chain of characters between two blanks. The WER then simply takes into account three types of errors: substitutions (S), insertions (I) and deletions (D).

- *Substitution (S)*: in a given word chain, one transcribed

word was different from the reference word.

- *Insertion (I)*: in a given word chain, a transcribed word was inserted with respect of the reference. The hypothesis counts one word more than the reference.
- *Deletion (D)*: in a given word chain, a word in the reference was not transcribed. The hypothesis counts one word less than the reference.

The following example sentences illustrates an alignment between a reference sentence (*Reference*) and an automatic transcription (*Hypothesis*) allowing the calculation of the WER:

| | | | | | | |
|---|---|---|---|---|---|---|
| **Reference** | How | are | you | | today | Patrick |
| | *S* | *D* | = | *I* | = | *S* |
| **Hypothesis** | Were | | you | here | today | playing |

Formally, the WER is calculated as follows:

$$WER = \frac{\#S + \#I + \#D}{\#reference\ words} \quad (1)$$

By definition, the WER therefore considers any type of error of equivalent importance. This is the main advantage of this metric: its simplicity of application and use. However, the WER does have limitations. Using the previous example, the word *Patrick* was transcribed as *playing*. An alternative transcription hypothesis could have been *Patricia*. In both cases, the WER would be identical to the reference, even though the nature of the error is different (*Patricia* is in the same grammatical category while *playing* is different from the reference word in terms of syntactics and semantics). Another limitation concerns the few categories considered (substitution, insertion, deletion) for the rate calculation carrying no additional information about the context.

### 2.2. Character Error Rate (CER)

The character error rate (CER) is based on the same principle as the WER but applied to character chains instead of word chains. It has already been used in the ASR domain [6]. Initially, it is particularly suitable for character-based languages such as Chinese or Japanese. For Latin languages, and in particular French, the CER allows, among other things, to give an indication of the nature of the errors: a low CER could indicate that the ASR system tends to generate words close to the reference (and thus potentially incorporating errors related to gender, number, tense, etc.) as opposed to a high CER, with transcription assumptions that are very distant from the references.

### 2.3. Part-of-speech Error Rate (POSER)

We also chose to use a metric allowing the calculation of the error rate on the part-of-speech (POS) classes of a transcription (POSER for *Part-of-speech Error Rate*). POSER allows us to know if the transcribed sentences are grammatically close to the reference ones, and to better characterize substitution errors. This rate is calculated with the same formula as the WER, except that POS are taken into account instead of words which relates to metadata of the transcribed words.

### 2.4. Lemma Error Rate (LER)

With a concept similar to the POSER and the WER, we did a Lemma Error Rate which consists of calculating the error rate of lemmas. We did two versions of this metric : one computing the WER and one computing the CER between the lemmas of the reference and the lemmas of the hypothesis.

### 2.5. Embeddings Error Rate (EmbER)

As previously exposed, the semantic aspect of a transcription is not taken into account in the WER metric. To address this, we consider a metric based on lexical word embeddings. Unlike existing metrics based on word embeddings, we aim at keeping the WER but weighting it: a word is no longer considered in a binary way (0 for a good transcript and 1 for an error), errors being weighted according to their semantic distance from the reference word. This distance is computed using the cosine similarity between the embeddings of the reference word and of the substituted transcribed word.

### 2.6. BERTScore

Developed for text generation [7], this metric aims at comparing a reference word and a hypothesis with respect to semantic proximity. The first step consists in obtaining the words and sub-words (*tokens*) of the reference and the hypothesis thanks to the WordPiece tokenizer used by BERT [8].

Then, given the sequence of contextualized embeddings of reference ($x1$,...,$xk$) and hypothesis ($\hat{x}1$,...,$\hat{x}m$), the cosine similarity is computed between each reference and hypothesis embeddings to obtain a score matrix weighted here with the inverse frequency of the document [7].

To compute the precision, we associate each token $x$ with a token $\hat{x}$ by selecting the token bringing the highest similarity. The recall is computed by associating each $\hat{x}$ token with an $x$ token in the same way. The f-measurement score, which we use in our experiments, is computed with the recall and the precision [7].

### 2.7. Sentence Semantic Distance (SemDist)

While previous metrics focus on words and characters, the principle of this metric [9] is to consider the complete sentence. In the ASR framework, the reference and the hypothesis are respectively transformed into their sentence embeddings using a SentenceBERT [10] model, *i.e.* a model of sentence embeddings using the contextual word embeddings of BERT [8]. It is then possible to compare these vectors with the cosine similarity. Our final measure is the average of the cosine similarities between each reference's sentence embeddings and its respective hypothesis.

## 3. Experimental protocol

In this section, we present the experimental protocol set up to apply the different metrics listed in Section 2. We describe the data used for our qualitative analysis of language model rescoring in Section 3.1, the ASR system and the POS tagger in Sections 3.2 and 3.3 respectively. Finally, we present the embeddings used by the different metrics and the lemmatizer.

### 3.1. Data

The French datasets used to train the ASR system are ESTER 1 and 2 [11, 12], EPAC [13], ETAPE [14], REPERE [15] and internal LIA data. Taken together, the corpora represent approximately 940 hours of audio of radio and television broadcast data. The evaluation of the systems is done on the REPERE test corpus, which is about 10 hours of audio data.

| | WER | CER | LER | LCER | dPOSER | uPOSER | EmbER | SemDist | BERTScore |
|---|---|---|---|---|---|---|---|---|---|
| **WER** | | | | | | | | | |
| **CER** | 89.34 | | | | | | | | |
| **LER** | 88.08 | 88.49 | | | | | | | |
| **LCER** | 87.10 | 98.31 | 91.40 | | | | | | |
| **dPOSER** | 92.96 | 90.02 | 92.70 | 89.51 | | | | | |
| **uPOSER** | 90.40 | 90.58 | 93.69 | 90.81 | 97.95 | | | | |
| **EmbER** | 96.51 | 91.51 | 86.57 | 88.78 | 91.00 | 88.98 | | | |
| **SemDist** | 71.81 | 64.78 | 62.22 | 62.60 | 65.33 | 64.13 | 75.73 | | |
| **BERTScore** | 74.63 | 74.27 | 72.60 | 73.00 | 74.09 | 74.25 | 84.51 | 63.35 | |

Table 1: *Averages of the Pearson correlations between the proposed metrics from both Base and Rescoring systems. For readability reasons, the values are multiplied by 100.*

### 3.2. Automatic Speech Recognition (ASR) system

The ASR system is based on an existing state-of-the-art recipe[1] that uses the Kaldi [16] toolkit. The acoustic model is a deep neural network based on the TDNNF [17] architecture. To make the system more robust to different acoustic conditions, the audio files were randomly perturbed in speed and volume (*i.e.* data augmentation) during the training process.

Three language models are used. The first is a trigram model trained with SRILM [18] and used directly by the ASR system. The second is a RNNLM, a deep neural network based language model, used in an a posterior rescoring process. The network consists of three TDNN layers interspersed with two LSTM layers. Also, a quadrigram model is used during the rescoring step. The training corpus and the vocabulary used to learn the trigram model, the RNNLM model and the quadrigram model are identical. The rescoring is optional as we want to observe its impact on the different metrics.

### 3.3. Tools

We used the POET tool[2][19], a POS tagger for French language based on Flair [20] contextual embeddings and used to automatically extract the morpho-syntactic information from words. We chose this labeler because it allows us to have both the generic classes of Universal Dependency (noun, adjective, adverb, etc.) but also a fine granularity thanks to additional information on these same labels (feminine plural noun, third person plural personal pronoun, etc.). We then propose two measures based on POS tags derived from the POSER (Section 2.3): one integrating the detailed classes (dPOSER) and one with the generic classes of Universal Dependency (uPOSER). Note that no manual POS tag annotation was used: both reference and hypothesis transcripts were automatically tagged. To obtain the lemmas, we used the Spacy lemmatizer for French[3].

For the EmbER metric (Section 2.5), we used Fasttext embeddings [21] and applied an error of 0.1 if the cosine similarity is above a threshold of 0.4, and 1 in other cases. The threshold was decided empirically given the cosine similarity between synonyms compare to cosine similarity between words randomly chosen.

For the SemDist metric (Section 2.7), the multilingual SentenceBERT embeddings was used. Finally, for the BERTScore, we use the default multilingual-BERT base model[4].

## 4. Experiments and Analysis

This section presents firstly an analysis of the six applied metrics presented in Section 2 in addition to the WER, and secondly a qualitative study of the impact of the language model rescoring process used in our ASR system.

### 4.1. Metrics analysis

In order to make a more in-depth analysis of our metrics, in particular to understand and estimate the links that they can maintain between them, we calculated a Pearson correlation between our different measurements for our two systems and averaged them in Table 1. The higher the score between two metrics, the more they are considered correlated. Clearly, the first remark is that not all metrics correlate with each other in the same way. SemDist is the metric that correlates the least with the others. This might be due to the fact that it is the only metric based on sentence embeddings in our experiments, going beyond the *word* dimension. This weak correlation implies that minimizing the WER would not correlate strongly with better performance on downstream tasks (*i.e.* extrinsic evaluation) using sentence embeddings. This idea is consistent with many publications in NLP and ASR that consider intrinsic ratings to be less relevant than extrinsic ratings [22, 23]. Indeed, the authors of SemDist [9] concluded that their metric correlated better with downstream tasks than the WER.

We can see that the metric that correlates best with BERTScore and SemDist is EmbER, all three of which are based on embeddings, while the metric that correlates best with EmbER is WER. This highlights that the Embedding Error Rate is a hybrid metric that has the advantage of correlating with WER and embeddings-based metrics.

An interesting observation to make is that LER correlate the best with uPOSER and has a better correlation with dPOSER than LCER. It seems that part-of-speech and lemmas share some similarity : if the lemma is wrong, the POS is often wrong. Also, the LCER and the CER have a correlation of 0.9831 which probably means that when the CER is high, there is a good chance that the word is wrong too and so is the lemma. On the other hand, it also means that the LCER does not bring more information than the CER.

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

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

[3] https://github.com/explosion/spacy-models/releases/tag/fr_core_news_md-3.2.0

[4] https://github.com/Tiiiger/bert_score

| System | WER | CER | dPOSER | uPOSER | LER | LCER | SemDist | BERTScore | EmbER |
|---|---|---|---|---|---|---|---|---|---|
| **Base** | 15.45 | 8.57 | 14.59 | 12.22 | 14.35 | 8.78 | 7.89 | 9.12 | 12.33 |
| **Rescoring** | 13.24 | 7.70 | 12.51 | 10.79 | 12.08 | 8.00 | 7.18 | 8.38 | 10.79 |
| *Reduction* | -14.3 % | -10.2 % | -14.3 % | -11.7 % | -15.8 % | -8.8 % | -9.0 % | -8.1 % | -12.5 % |

Table 2: *Performance comparison of the Base and Rescoring systems using different metrics. The observed reduction between the two systems, in relative value, is also provided.*

### 4.2. Rescoring Impact

In order to improve the performance of the ASR, rescoring was acheived using a RNNLM, a deep neural network based language model.

Table 2 presents the results obtained with the different metrics applied to the automatic transcriptions from the ASR system without (Base) and with hypothesis reordering (Rescoring). As expected, rescoring improves the results with a decrease of error rates independently of the used metric: an improvement is thus visible at the level of words, characters, syntax and semantics. The gains for each metric are also provided in Table 2. They mainly highlight the fact that the relative gain obtained on the WER is the highest compared to the other metrics. Depending on the purpose of the system, the quality of a transcription can be defined by its grammatical, lexical or semantic similarity with the reference. We therefore imagine that the benefits obtained thanks to this rescoring step are not as significant as what the WER suggests. In comparison, the SemDist and BERTScore metrics have the lowest relative gains, which tends to make us say that rescoring only partially corrects transcribed words that were semantically far from their reference. The proposed EmbER, which is a mixed measure between WER and embeddings, seems to take into account the syntactic and semantic level, with a gain between that of WER and embedding measurements. Overall, language model rescoring contributes less to the improvement of the semantic level (SemDist, BERTScore and EmbER) compared to the syntactic level, visible with a huger reduction on the character, POS and lemma based measures.

Thanks to the meta information annotated in the REPERE corpus, we could observe that the rescoring process deteriorates performances on utterances of spontaneous speech. In average, utterances presenting more errors after the rescoring step contained 1.23 times more spontaneity information (elisions, reduction, truncations and others disfluences).

This is in line with the hypothesis we made: a speech with too much disfluences (and so a mismatch between linguistic training and testing conditions) might be negatively impacted by rescoring.

With respect to POS, we propose in Table 3 to measure the average cosine distance between every reference and hypothesis word. We computed this distance without (Base) and with rescoring, while providing the relative reduction for each POS. This highlighted that Interjections (INTJ) and subordinating conjunctions (CCONJ), and to a lesser extent verbs (VERB) and adjectives (ADJ), are the word categories that benefit the most from rescoring while numbers (NUM) or determinants (DET) are among the POS classes that benefit the less from this additional step. The reason for the improvement of the interjections is probably because this POS is the one with the highest error rate.

| | Base | Rescoring | Reduction |
|---|---|---|---|
| **INTJ** | 14.07 | 10.45 | -3.63 |
| **CCONJ** | 9.83 | 6.82 | -3.01 |
| **VERB** | 6.10 | 4.20 | -1.90 |
| **ADJ** | 5.08 | 3.41 | -1.67 |
| **AUX** | 4.66 | 3.27 | -1.39 |
| **PRON** | 5.37 | 4.12 | -1.25 |
| **SCONJ** | 3.51 | 2.43 | -1.08 |
| **PROPN** | 6.72 | 5.82 | -0.90 |
| **NOUN** | 3.34 | 2.57 | -0.77 |
| **ADV** | 3.23 | 2.49 | -0.74 |
| **ADP** | 2.90 | 2.25 | -0.65 |
| **DET** | 2.95 | 2.42 | -0.53 |
| **NUM** | 2.96 | 2.62 | -0.34 |

Table 3: *Average semantic distance per POS between each word from the reference and their associated word from the hypothesis. For readability. the values are multiplied by 100.*

## 5. Conclusions and Perspectives

In this study, we applied different measures in addition to the WER metric to ASR systems in order to reveal different linguistic dimensions (grammatical, semantic, etc.) to transcription errors.

We have chosen to verify their relevance by studying the impact of a posteriori hypothesis reordering on ASR systems using language models. Our study showed that the gains are not equivalent depending on the metric considered, thus highlighting the limitations of WER alone to study improvements at the lexical, grammatical or semantic level. It is important to note that the rescoring improve overall performances, though the increase in performance is not always visible locally.

In the continuity of this work, we would like to extend this analysis by combining the measures. Indeed, we have been interested here in these metrics independently, but it seems relevant to study, for example, semantic measures on identified POS (*e.g.*, compare BERTScore on personal names and adjectives). Also, this study focuses on the linguistic aspect of ASR, while we observed that segments with high speech spontaneity clues may be negatively impacted by the rescoring process. It would then be interesting to continue this study at the acoustic level, by looking into other audio factors such as noise or speech overlap. In the longer term, it would be interesting to evaluate the correlation between our metrics and human perception of errors.

## 6. Acknowledgments

This work was supported by the DIETS project financed by the Agence Nationale de la Recherche (ANR) under contract ANR-20-CE23-0005. It was granted access to the HPC resources of IDRIS under the allocation 2021-A0111012991 made by GENCI.

# 7. References


[1] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 8599–8603.

[2] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*. PMLR, 2016, pp. 173–182.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[4] M. Morchid, R. Dufour, and G. Linarès, "Impact of word error rate on theme identification task of highly imperfect human–human conversations," *Computer Speech & Language*, vol. 38, pp. 68–85, 2016.

[5] S. Mdhaffar, Y. Estève, N. Hernandez, A. Laurent, R. Dufour, and S. Quiniou, "Qualitative evaluation of asr adaptation in a lecture context: Application to the pastel corpus." in *InterSpeech*, 2019, pp. 569–573.

[6] M. Xu, S. Li, and X.-L. Zhang, "Transformer-based end-to-end speech recognition with local dense synthesizer attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5899–5903.

[7] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[9] S. Kim, D. Le, W. Zheng, T. Singh, A. Arora, X. Zhai, C. Fuegen, O. Kalinli, and M. L. Seltzer, "Evaluating user perception of speech recognition system quality with semantic distance metric," *arXiv preprint arXiv:2110.05376*, 2021.

[10] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[11] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri, "Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news." in *International Conference on Language Resources and Evaluation (LREC)*, 2006, pp. 139–142.

[12] S. Galliano, G. Gravier, and L. Chaubard, "The ester 2 evaluation campaign for the rich transcription of french radio broadcasts," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[13] Y. Esteve, T. Bazillon, J.-Y. Antoine, F. Béchet, and J. Farinas, "The epac corpus: manual and automatic annotations of conversational speech in french broadcast news," in *International Conference on Language Resources and Evaluation (LREC)*, 2010.

[14] G. Gravier, G. Adda, N. Paulsson, M. Carré, A. Giraudel, and O. Galibert, "The etape corpus for the evaluation of speech-based tv content processing in the french language," in *International Conference on Language Resources and Evaluation (LREC)*, 2012, pp. 114–118.

[15] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The repere corpus: a multimodal corpus for person recognition," in *International Conference on Language Resources and Evaluation (LREC)*, 2012, pp. 1102–1107.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE Signal Processing Society, 2011.

[17] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Interspeech*, 2018, pp. 3743–3747.

[18] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.

[19] Y. Labrak and R. Dufour, "Antilles: An open french linguistically enriched part-of-speech corpus," in *International Conference on Text, Speech, and Dialogue*. Springer, 2022, pp. 28–38.

[20] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in *Proceedings of the 27th international conference on computational linguistics*, 2018, pp. 1638–1649.

[21] P. Bojanowski, É. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[22] Y.-Y. Wang, A. Acero, and C. Chelba, "Is word error rate a good indicator for spoken language understanding accuracy," in *IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2003, pp. 577–582.

[23] G. Glavaš, R. Litschko, S. Ruder, and I. Vulić, "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*, 2019, pp. 710–721.