Approches cascade et de bout-en-bout pour la traduction automatique de la parole en pictogrammes

Cécile Macaire, Chloé Dion, Didier Schwab, Benjamin Lecouteux, Emmanuelle Esperança-Rodier

Approches cascade et de bout-en-bout pour la traduction automatique de la parole en pictogrammes

Cécile Macaire[1] Chloé Dion[1] Didier Schwab[1] Benjamin Lecouteux[1]
Emmanuelle Esperança-Rodier[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France
prenom.nom@univ-grenoble-alpes.fr

RÉSUMÉ

La traduction automatique de la parole en pictogrammes (Parole-à-Pictos) est une nouvelle tâche du Traitement Automatique des Langues (TAL) ayant pour but de proposer une séquence de pictogrammes à partir d'un énoncé oral. Cet article explore deux approches distinctes : (1) en cascade, qui combine un système de reconnaissance vocale avec un système de traduction, et (2) de bout-en-bout, qui adapte un système de traduction automatique de la parole. Nous comparons différentes architectures état de l'art entraînées sur nos propres données alignées parole-pictogrammes. Nous présentons une première évaluation automatique des systèmes et réalisons une évaluation humaine pour analyser leur comportement et leur impact sur la traduction en pictogrammes. Les résultats obtenus mettent en évidence la capacité d'une approche en cascade à générer des traductions acceptables à partir de la parole lue et dans des contextes de la vie quotidienne.

ABSTRACT

Cascade and End-to-End Approaches for Automatic Speech-to-Pictograms Translation

The automatic translation of speech into pictograms (Speech-to-Pictograms) is a new Natural Language Processing (NLP) task whose purpose is to generate a sequence of pictograms based on a speech utterance. This article explores two distinct approaches : (1) the cascade approach, which combines a speech recognition system with a machine translation system, and (2) the end-to-end approach, which adapts a speech translation system. We compare different state-of-the-art architectures trained on our own aligned speech-to-pictogram data. We present a first automatic evaluation of the systems and conduct a human evaluation to analyze their behavior and their impact on pictogram translation. The results highlight the ability of the cascade approach to generate acceptable translations from spoken language in everyday life situations.

MOTS-CLÉS : Pictogrammes, Parole, Traduction Automatique.

KEYWORDS: Pictograms, Speech, Machine Translation.

1 Introduction

La Communication Alternative et Augmentée (CAA) regroupe un ensemble d'outils et de stratégies conçus pour faciliter la communication des individus confrontés à des troubles du langage (Beukelman & Mirenda, 2017). Ces troubles impactent un ensemble de capacités langagières, allant de la production et la compréhension de la parole à l'écoute, la lecture et l'écriture. Ils peuvent avoir différentes origines, telles que certaines maladies génétiques, des troubles du spectre autistique, un déficit

intellectuel, pour en citer quelques-uns. Nous retrouvons, dans la CAA, l'utilisation de pictogrammes, pour transmettre des messages dans des situations de la vie quotidienne. Un pictogramme est une représentation graphique associée à un concept (objet, personne, action, etc.) (Pereira *et al.*, 2022b). Ceux-ci présentent plusieurs avantages, notamment, qu'ils permettent de visualiser la syntaxe, de manipuler des mots et de faciliter l'accès au langage (Cataix-Nègre, 2017). D'un point de vue social, une enquête menée par la Croix-Rouge (2021) a identifié une réduction du stress, une augmentation de l'autonomie et un impact positif du bien-être général des utilisateurs de CAA.

Pourtant, selon la même étude, la CAA est confrontée à différents freins environnementaux qui limitent son utilisation et sa diffusion. L'étude cite spécifiquement le manque de sensibilisation des accompagnants au potentiel de la CAA et la difficulté d'accéder aux outils (absence d'informations, de formation, de moyens financiers et de temps).

Nous pensons que la mise en place de systèmes de traduction de la parole en une séquence de pictogrammes, tâche que nous appellerons Parole-à-Pictos (PAP), pourrait permettre de relever ces défis. Un système PAP prédit une suite de termes, chacun associé à un pictogramme unique ARASAAC [1] à partir d'un segment audio (cf. Figure 1). Notre objectif est de construire le premier système PAP pour le français.

FIGURE 1 – Illustration de la tâche PAP avec la traduction d'un segment audio en pictogrammes ARASAAC.

Pour cela, nous proposons deux approches. La première s'appuie sur un système en cascade, qui imbrique un modèle de Reconnaissance Automatique de la Parole (RAP) avec un modèle de Traduction Automatique (TA). La deuxième approche adapte une architecture de bout-en-bout de Traduction de la Parole (TP) pour la tâche Parole-à-Pictos. Nous présentons une première évaluation automatique et humaine sur nos propres jeux de données. Nous résumons nos contributions ci-dessous :

- La présentation de deux approches pour traduire automatiquement la parole en une séquence de pictogrammes.
- La construction de trois corpus de données alignées parole-texte-pictogrammes pour cette tâche [2].
- L'implémentation et la publication de plusieurs modèles de Reconnaissance de la Parole, de Traduction Automatique et de Traduction de la parole affinés sur ces jeux de données. Le code et les modèles sont disponibles en ligne [3] et les expériences sont entièrement reproductibles.
- La présentation d'une première évaluation automatique et humaine des modèles pour la tâche Parole-à-Pictos.

1. https://arasaac.org/
2. https://www.ortolang.fr/market/corpora/propicto
3. https://github.com/macairececile/speech-to-pictograms

2 État de l'art

Les précédents travaux se sont majoritairement focalisés sur la traduction du texte vers les pictogrammes, plutôt qu'à partir de la parole. Sevens *et al.* (2015) ont notamment proposé Text2Picto, un système de traduction texte-pictogrammes pour le néerlandais, ensuite étendu à l'anglais, à l'espagnol (Sevens, 2018) et au français (Norré *et al.*, 2021). De récents travaux (Pereira *et al.*, 2022a, 2023) se sont intéressés à la prédiction d'un pictogramme selon le contexte, en utilisant des modèles de type BERT (Devlin *et al.*, 2019). Le but est de proposer un modèle prédictif pour compléter une phrase en construction dans les systèmes de CAA.

La traduction automatique de la parole en une séquence de pictogrammes du Français est étudiée pour la première fois dans Vaschalde *et al.* (2018). Leur méthodologie s'appuie sur l'adaptation du système Text2Picto (Vandeghinste *et al.*, 2017) à la parole. Le modèle proposé imbrique quatre modules : un système de Reconnaissance Automatique de la Parole (RAP), un système de simplification, un modèle de désambiguïsation lexicale et un dernier module qui affiche la séquence de pictogrammes. L'évaluation est réalisée sur deux jeux de données. Le premier regroupe 15 histoires pour enfants manuellement traduites en pictogrammes. Le second est un ensemble de 20 phrases extraites du corpus ESLO (Baude & Dugua, 2017). Aucune évaluation automatique ou humaine ne sont rapportées. Récemment, les travaux de Macaire *et al.* (2022, 2023) ont exploré la traduction Parole-à-Pictos en proposant Voice2Picto. Bien que le système propose une approche novatrice, il ne compare pas différents systèmes de RAP. De plus, la traduction utilise un vocabulaire pictographique non adapté à la parole et aucune donnée spécifique à cette tâche n'est employée. Aucune évaluation, qu'elle soit automatique ou humaine, n'a été réalisée.

Pour la traduction de la parole en pictogrammes, nous nous appuyons sur les travaux précédents qui associent un système de Reconnaissance Automatique de la Parole et un système de Traduction Automatique.

Reconnaissance Automatique de la Parole `Wav2Vec2.0` (Baevski *et al.*, 2020) est un modèle basé sur l'apprentissage auto-supervisé. Celui-ci apprend des représentations robustes de la parole sur une collection importante de données non étiquetées pendant la phase dite de pré-entraînement. L'architecture est ensuite affinée sur un jeu de données étiqueté pour une tâche en aval. Plus récemment, deux modèles multimodaux et multilingues montrent des résultats compétitifs sans nécessiter une phase d'affinage. `Whisper` (Radford *et al.*, 2023) utilise l'architecture encodeur-décodeur Transformer (Vaswani *et al.*, 2017). Le modèle est appris sur 680 000 heures de données étiquetées multilingues (plus de 100 langues). `SeamlessM4T` (Barrault *et al.*, 2023) est un modèle massif de traduction automatique multimodale (traduction parole-parole, parole-texte, texte-parole, texte-texte et transcription) et multilingue sur une centaine de langues. Contrairement à `Whisper`, `SeamlessM4T` préserve les éléments de la prosodie et du style vocal dans toutes les langues couvertes.

Traduction Automatique Ott *et al.* (2018) présente un modèle neuronal Transformer (Vaswani *et al.*, 2017) séquence-à-séquence. L'architecture utilise un vocabulaire commun à chaque paire de langues. Les données sont tokenisées en sous-mots avec l'algorithme *Byte-Pair Encoding*. L'architecture est entraînée à partir de zéro. Les modèles suivants sont pré-entraînés sur des données multilingues. Liu *et al.* (2020) présente `mBART`, un modèle auto-encodeur séquence-à-séquence pré-entraîné sur une quantité importante de données monolingues dans plusieurs langues. L'architecture applique un objectif BART (Lewis *et al.*, 2020), modèle de type Transformer. L'article souligne l'avantage de `mBART` sur des langues ne figurant pas dans les données de pré-entraînement. Raffel *et al.* (2020) propose `T5`, une approche basée sur l'apprentissage par transfert. Chaque donnée textuelle

est considérée, en entrée, comme un problème texte-à-texte, permettant ainsi de réaliser différentes tâches (résumé de documents, analyse de sentiments, traduction automatique, etc.) via un modèle unique. Le modèle utilise 20TB de données textuelles de langues anglaise, française, roumaine et allemande. Costa-jussà *et al.* (2022) propose `NLLB`, un modèle de type Transformer massivement multilingue capable de traduire automatiquement dans 200 langues. Cette couverture linguistique peut être bénéfique entre deux langues apparentées via un transfert interlinguistique (Conneau *et al.*, 2020; Fan *et al.*, 2021). Plusieurs travaux présentent des approches basées sur les représentations de phrases, notamment LASER(Artetxe & Schwenk, 2019), LabSE (Feng *et al.*, 2022) et SONAR (Duquenne *et al.*, 2023). Ce dernier obtient des résultats compétitifs en TA par rapport au modèle NLLB 1B.

Traduction Automatique de la Parole La traduction automatique de parole de bout-en-bout est explorée dans plusieurs travaux. Nous pouvons citer `Fairseq S2T` (Wang *et al.*, 2020), qui combine un modèle RNN et Transformer. Ye *et al.* (2021) présentent `XSTNET`, un modèle transversal parole-texte avec `Wav2Vec2.0` comme encodeur vocal, suivi d'un entraînement progressif multitâche (modèle de TA pré-entraîné et affinage multitâche). Plus récemment, les travaux de Ye *et al.* (2022) proposent `ConST`, un modèle fondé sur une approche d'apprentissage contrastive. Celui-ci cherche à encoder les représentations audio et textuelles similaires dans un espace proche. Composé de quatre modules, `ConST` intègre un encodeur vocal utilisant les représentations `Wav2vec2.0`, une couche de plongement de mots et un encodeur-decodeur Transformer. Les scores BLEU rapportés sur MUST-C (Di Gangi *et al.*, 2019) démontrent des performances état de l'art, notamment pour des paires de langues peu dotées.

3 Méthode proposée

Nous appliquons deux approches pour la tâche Parole-à-Pictos. La première est une approche cascade constituée d'un système de RAP et d'un système de TA. Ici, la transcription fournie par le système de RAP est le point d'entrée du système de TA, dont le but est de traduire la langue source (ici le français) dans la langue cible. Pour la tâche Parole-à-Pictos, la langue cible est ce que nous nommons "langage pictographique" qui correspond à la séquence de termes (mot unique, expression polylexicale, ou phrase entière), chacun associé à un pictogramme ARASAAC. Pour les systèmes de RAP, nous comparons `Wav2Vec2.0`, `Whisper` et `SeamlessM4T`. L'objectif est de confronter les performances d'un modèle ajusté à nos données à celles de modèles massivement multilingues, multitâches et du domaine général, mais non reproductibles. Nous considérons, pour la Traduction Automatique, les modèles état de l'art présentés Section 2. Notre objectif est de comparer une architecture à entraîner à partir de zéro (*from scratch*) avec des architectures pré-entraînées sur des données multilingues et à affiner sur ces propres données. La seconde approche adapte les systèmes de bout-en-bout de Traduction Automatique de la parole à notre tâche. Dans cet article, nous testons le modèle `ConST`.

4 Données

Nous construisons deux ensembles de données issus de deux corpus de parole préexistants pour entraîner nos systèmes. Ces deux corpus se différencient par le type de parole qu'ils contiennent

(parole lue et parole spontanée). L'objectif est d'évaluer la robustesse des modèles face à diverses situations acoustiques. Un jeu de données supplémentaire est également utilisé pour l'évaluation, qui se rapproche des interactions du public cible.

Les pictogrammes utilisés proviennent d'ARASAAC, une ressource riche de plus de 25 000 pictogrammes uniques, continuellement mise à jour. Distribués sous la licence Creative Commons CC-BY-NC-SA et téléchargeables gratuitement, ces pictogrammes sont très largement utilisés dans la communauté CAA.

Propicto-orféo Nous récupérons les données alignées parole/texte issues du Corpus d'Étude pour le Français Contemporain (CEFC) (Benzitoun *et al.*, 2016), comprenant un ensemble de 12 corpus sources. Nous retrouvons des situations de parole diverses (dialogues, réunions, etc.) et dans des domaines variés. Propicto-orféo contient 290 036 segments audio pour un total de 233 h. Chaque segment audio est accompagné d'une transcription. À partir de celles-ci, nous appliquons la méthode présentée par Macaire *et al.* (2024) pour générer une traduction en pictogrammes, qui suit des règles et un lexique précis. Les données ont la forme suivante, avec *tokens* se référant à la liste des termes associés à chaque pictogramme présent dans *pictos* :

```
{
  "id": "cefc-tcof-Hen_sai_vin_reunion_08-190",
  "text": "ça fera trop rapproché",
  "pictos": [9829, 6906, 6190, 25708, 6879],
  "tokens": "prochain celle-là faire trop approcher"
}
```

Propicto-commonvoice Nous récupérons la partie française du corpus de parole lue CommonVoice version 15 (Ardila *et al.*, 2020). Cette version comprend 967 heures d'enregistrements issues de 17 911 locuteurs uniques. Sur le même principe décrit précédemment, nous appliquons la méthode de Macaire *et al.* (2024) pour générer la traduction de chaque segment audio en pictogrammes.

Propicto-eval Nous utilisons un jeu de données test pour évaluer les différentes approches sur un domaine et un type de parole restreints. Propicto-eval est un corpus de parole lue multilocuteurs (62 au total). Les données textuelles proviennent d'histoires pour enfants, de situations de la vie quotidienne et de phrases du domaine médical. Ces contextes sont particulièrement pertinents, car ils reflètent les types d'interactions de notre public cible.

5 Expériences

Données et pré-traitement Nous répartissons les données Propicto-orféo et Propicto-commonvoice en trois ensembles entraînement, validation et test selon une répartition 90/5/5. Nous supprimons la ponctuation et convertissons les transcriptions en minuscules. Chaque segment audio représente une phrase de moins de 30 secondes, la taille maximale pouvant être encodée par les systèmes de RAP. Les segments ne comprenant pas de traductions en pictogrammes ne sont pas conservés dans notre ensemble. Ces segments contiennent des disfluences ou des termes non traduits en pictogramme, conséquence de la limite de la méthode de Macaire *et al.* (2024) et des limites d'ARASAAC (certains domaines sont sous-représentés en pictogrammes). Les données sont détaillées Table 1.

	Propicto-commonvoice		Propicto-orféo	
	# phrases	# heures	# phrases	# heures
entraînement	527 554	756	231 374	147
validation	16 132	25	28 796	18
test	16 132	26	29 009	14

TABLE 1 – Répartition des données en trois ensembles (entraînement, validation, test) avec, pour chaque corpus, le nombre de phrases et le nombre d'heures.

Détails des entraînements Nous utilisons la boîte à outils SpeechBrain (Ravanelli *et al.*, 2021) et la recette fournie [4] pour affiner le modèle `Wav2Vec2.0` de RAP, avec, comme modèle pré-entrainé, *LeBenchmark/wav2vec2-FR-7K-large* (Evain *et al.*, 2021). Les segments audio de moins de 3 secondes et de plus de 10 secondes ont été écartés de l'entraînement pour éviter les segments audio trop courts ou vides.

Pour les différents systèmes de Traduction Automatique (TA), nous exploitons deux boîtes à outils : Fairseq (Ott *et al.*, 2019) et HuggingFace (Wolf *et al.*, 2020). Nous adaptons la recette proposée par Fairseq [5] du modèle de traduction *from scratch* `NMT`. Une phase de tokenisation (BPE) segmente le texte en unités de sous-mots. Un vocabulaire de 10 000 jetons est généré. Fairseq est également utilisée pour affiner le modèle `mBART`, ici, *mbart-large-cc25* appris sur 25 langues [6]. La même méthode de tokenisation décrite précédemment est appliquée. L'affinage des modèles `T5-large` et `NLLB-200` (*facebook/nllb-200-1.3B*) est réalisée en adaptant la recette proposée par HuggingFace [7]. Les principaux paramètres des modèles sont décrits Table 2 [8].

Modèle ↓	# paramètres	taux d'apprentissage	taille du lot	# epoch
Whisper large-v3 (Radford *et al.*, 2023)	1550M	-	-	-
SeamlessM4T-Large v2 (Barrault *et al.*, 2023)	2.3B	-	-	-
Wav2Vec2 (Baevski *et al.*, 2020) + CTC greedy search	318,7M	1e-4	8	30
NMT (Ott *et al.*, 2018)	51M	5e-4	8	40
mBART25 (Liu *et al.*, 2020)	610M	3e-5	8	40
T5-large (Raffel *et al.*, 2020)	220M	2e-5	32	40
NLLB-200 (Costa-jussà *et al.*, 2022)	600M	2e-5	32	40
ConST (Ye *et al.*, 2022)	150M	1e-4	8	40

TABLE 2 – Paramètres des modèles de Reconnaissance Automatique de la Parole, de Traduction Automatique et de Traduction Automatique de la Parole.

Enfin, nous suivons le pipeline basé sur Fairseq [9] pour entraîner le modèle `ConST`. Le modèle pré-entrainé *LeBenchmark/wav2vec2-FR-7K-base* est employé.

4. https://github.com/speechbrain/speechbrain/tree/develop/recipes/LibriSpeech/ASR/CTC
5. https://github.com/facebookresearch/fairseq/blob/main/examples/translation/
6. https://github.com/facebookresearch/fairseq/tree/main/examples/mbart
7. https://huggingface.co/docs/transformers/tasks/translation#train
8. Veuillez vous référer aux recettes citées pour les informations complètes sur les paramètres, ceux-ci n'ont pas été modifiés.
9. https://github.com/ReneeYe/ConST/tree/main

Résultats Nous présentons les résultats des trois approches de Reconnaissance Automatique de la Parole dans la Table 3. Notre évaluation rapporte le taux d'erreur au niveau des mots (Woodard & Nelson, 1982; Morris *et al.*, 2004). Pour les deux corpus, l'approche `Wav2Vec2.0` obtient les meilleures performances. Sur Propicto-commonvoice, la différence entre les trois modèles est minime (3 points d'écart), contrairement à Propicto-orféo avec 18,9 points séparant `Wav2vec2.0` et `Whisper`. Une hypothèse plausible pour expliquer ce phénomène pourrait résider dans le type de parole que renferment les corpus. Propicto-orféo est un corpus de parole spontanée contenant des chevauchements entre locuteurs et des disfluences (hésitations, répétitions, faux-départ). Nous supposons que `Whisper` et `SeamlessM4T`, ayant été entraînés sur une majorité de parole lue, se généralisent donc mal à des corpus plus complexes.

Modèle ↓	validation		test	
	Propicto-commonvoice	Propicto-orféo	Propicto-commonvoice	Propicto-orféo
Whisper large-v3 (Radford *et al.*, 2023)	-	-	14.34	37.69
SeamlessM4T-Large v2 (Barrault *et al.*, 2023)	-	-	12.45	46.50
Wav2Vec2 + CTC greedy search (Baevski *et al.*, 2020)	9.14	23.24	**11.21**	**27.56**

TABLE 3 – Taux d'erreur au niveau des mots (%) rapportés sur Propicto-commonvoice et Propicto-orféo entre les trois modèles de RAP.

Les résultats des modèles de traduction texte-à-pictogrammes sont présentés Table 4. Nous rapportons le score BLEU [10] par modèle et par corpus. Celui-ci est calculé en comparant la séquence de termes prédits par rapport à la séquence de termes "gold" (*tgt* : "prochain celle-là faire trop approcher", *hyp* : "celle-là faire non trop approcher"). Nous constatons des scores similaires entre les deux corpus. `mBART` présente un écart significatif avec les autres modèles (plus de 12 points en moins). De plus, le modèle `NMT`, approche entraînée à partir de zéro surpasse `mBART`. Les résultats ne démontrent pas un apport important des modèles pré-entraînés multilingues pour cette tâche de traduction. Nous n'excluons pas que la méthode de segmentation utilisée peut influencer les performances. D'autres techniques sont à explorer.

Modèle ↓	validation		test	
	Propicto-commonvoice	Propicto-orféo	Propicto-commonvoice	Propicto-orféo
Neural Machine Translation (NMT) (Ott *et al.*, 2018)	86.06	**87.28**	82.60	**87.43**
mBART25 (Liu *et al.*, 2020)	72.39	75.26	72.31	75.62
T5-large (Raffel *et al.*, 2020)	86.36	85.21	86.58	85.88
NLLB-200 (Costa-jussà *et al.*, 2022)	**87.41**	86.32	**87.66**	86.92

TABLE 4 – Scores BLEU des quatre modèles de Traduction Automatique par corpus. Les résultats sont présentés sur les données de validation et de test.

C'est lors de l'association des systèmes de Reconnaissance Automatique de la Parole et les systèmes de Traduction Automatique (notre approche cascade) que certains modèles se démarquent par rapport aux autres. La Table 5 présente les scores BLEU sur les données test en combinant chaque modèle. Pour Propicto-commonvoice, l'association de `SeamlessM4T` et `NLLB-200` obtient le score BLEU le plus élevé. Quant aux deux autres modèles de RAP avec `NLLB-200`, leurs scores sont très étroitement alignés, avec une différence de seulement 0,75. Nous expliquons cette similarité dans les

10. Les modèles sont évalués avec sacreBLEU (Post, 2018).

scores par le fait qu'il y a un écart non significatif entre les performances des systèmes de RAP. Les performances de Propicto-orféo subissent une baisse significative lorsque le système de traduction utilise les transcriptions prédites par le système de RAP en entrée. Précisément, nous observons une diminution de plus de 24 points du score BLEU, pour atteindre un score de 62,48 avec l'association de `Wav2Vec2.0` et `NLLB-200`. Le système de RAP impacte fortement les performances de la traduction en pictogrammes. De plus, bien que le modèle `NMT` soit le plus performant en TA, c'est `NLLB-200` et `T5-large` avec `Wav2Vec2.0` qui obtiennent les meilleurs scores BLEU dans notre approche cascade. Nous supposons que les modèles pré-entraînés massivement multilingues sont plus robustes lorsqu'ils sont confrontés à des termes déformés par le système de RAP.

Modèle RAP ↓	Modèle TA ↓	test	
		Propicto-commonvoice	Propicto-orféo
Whisper large-v3	NMT	73.72	58.07
	mBART25	67.02	52.05
	T5-large	78.67	57.80
	NLLB-200	**79.49**	58.82
SeamlessM4T-Large v2	NMT	73.72	52.38
	mBART25	67.61	48.71
	T5-large	79.45	53.96
	NLLB-200	**80.15**	54.86
Wav2Vec2 + CTC greedy search	NMT	73.45	61.37
	mBART25	67.02	55.49
	T5-large	78.55	**61.66**
	NLLB-200	**79.49**	**62.48**

TABLE 5 – Scores BLEU sur les données test obtenus en combinant chaque modèle de RAP avec les modèles de TA.

Nous concluons nos expériences en présentant le score BLEU Table 6 du modèle de traduction de la parole de bout-en-bout `ConST`. Les résultats indiquent des performances inférieures à celles d'une approche en cascade. Cependant, nous observons des résultats compétitifs, voire meilleurs par rapport à certaines associations de modèles RAP et TA. D'autres architectures restent à explorer dans des travaux futurs.

Modèle ↓	validation		test	
	Propicto-commonvoice	Propicto-orféo	Propicto-commonvoice	Propicto-orféo
ConST (Ye *et al.*, 2022)	73.13	62.21	71.65	60.21

TABLE 6 – Scores BLEU obtenus sur les données de validation et test par corpus du modèle de traduction de la parole de bout-en-bout `ConST`.

Le score BLEU n'offre pas d'informations précises sur les comportements spécifiques de chaque approche dans le contexte d'une traduction en pictogrammes. À cette fin, nous conduisons une évaluation humaine.

6 Évaluation humaine

Nous menons une évaluation humaine sur les deux modèles ayant obtenu le score BLEU le plus élevé pour chaque corpus. Nous adaptons un cadre analytique de conseils et de procédures pour mesurer la qualité d'une traduction, défini par Burchardt (2013), MQM [11]. Cette évaluation permet de déterminer si la traduction proposée répond aux spécifications convenues par les parties prenantes. L'évaluation analytique associe les erreurs à des mots et à des phrases spécifiques du texte (source et/ou cible). Le rôle de l'évaluateur est d'examiner le texte traduit par rapport au texte source et aux spécifications, puis d'annoter les erreurs conformément à la métrique (c'est-à-dire identifier, marquer et attribuer un type d'erreur et un niveau de gravité).

En analysant de façon globale les traductions proposées par les différents systèmes, nous sélectionnons 12 erreurs réparties en 4 catégories :

— *Précision* — ajout, omission, erreur de traduction, sur-traduction, sous-traduction,
— *Fluidité* — inintelligible, ambigu, cohésion, ordre des mots, offensif,
— *Vérité* — exhaustivité (le texte source est-il en adéquation avec le public cible ?),
— *Design* — longueur (écart important entre la longueur du texte source et celle du texte cible).

À chaque erreur est associé un niveau de gravité :

— *neutre*,
— *mineur* (aucune incidence sur la facilité d'utilisation ou la compréhensibilité du contenu),
— *majeur* (incidence sur la facilité d'utilisation sans pour autant rendre la traduction inutilisable),
— *critique* (traduction inutilisable, ce qui entraîne des dommages aux personnes, au matériel ou à la réputation d'une organisation si celles-ci ne sont pas corrigées).

Deux annotateurs experts du projet ont annoté 100 phrases sélectionnées aléatoirement parmi les données tests de Propicto-orféo et Propicto-commonvoice. Nous présentons Table 7 le Score de Qualité Globale par modèle. Celui-ci est calculé en multipliant la note de pénalité (qui résulte de l'association du nombre d'erreurs par catégorie et le niveau de gravité auquel est associé un poids - neutre : 0, mineur : 1, majeur : 5, critique : 25) par la valeur maximale (généralement 100). Le système de traduction ne peut être validé si le Score de Qualité Globale (SQG) est inférieur à la valeur de seuil. Le SQG est calculé en divisant le nombre total de pénalités (nombre d'erreurs multiplié par le poids du niveau de gravité associé à chacune) avec le nombre total de termes évalués. Par les différentes observations, les parties prenantes ont défini la limite de compréhension et d'utilisation d'une traduction à deux erreurs majeures et une erreur mineure, ce qui équivaut à une valeur de seuil de 89 ($100 - ((2 * 5) + (1 * 1))$).

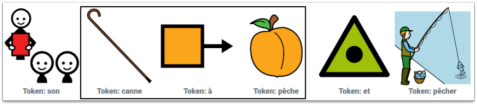
Modèle ↓		Score de Qualité Globale
Propicto-orféo	Wav2Vec2 + CTC / T5-large	45.78
	Wav2Vec2 + CTC / NLLB-200	44.56
Propicto-commonvoice	Wav2Vec2 + CTC / NLLB-200	80.65
	SeamlessM4T-Large v2 / NLLB-200	81.96
Propicto-eval	Whisper large-v3 / T5-large	87.78
	Whisper large-v3 / NLLB-200	**90.01**

TABLE 7 – Score de qualité globale calculé sur 100 phrases annotées pour les deux meilleurs modèles par corpus.

11. https://themqm.org/

Les systèmes de traduction sur Propicto-orféo et Propicto-commonvoice n'ont pas atteint un score supérieur au seuil, rejetant ainsi leur utilisation auprès d'un public cible. Ce score s'explique par certains comportements observés par les annotateurs :

- la traduction des entités nommées est inexacte, par exemple une ville sera traduite par un pictogramme représentant une personne et inversement,
- les termes polylexicaux sont découpés en n pictogrammes au lieu d'un, par exemple "canne à pêche",
- certains homonymes sont incorrects ("avocat" pour le fruit au lieu du métier), ce problème découle des données utilisées,
- certains termes ne sont pas traduits, majoritairement dû aux erreurs générées par les systèmes de RAP,
- les chiffres sont découpés en plusieurs pictogrammes (800 traduit par 8 et 100).



Nous réalisons une dernière évaluation sur l'ensemble test Propicto-eval. En appliquant les différentes approches, c'est l'association de `Whisper` avec `T5-large` (score BLEU de 77,23) et `Whisper` avec `NLLB-200` (score BLEU de 74,97) qui obtiennent les meilleures performances. L'évaluation humaine conduite valide l'utilisabilité de notre approche avec `Whisper` et `NLLB-200` sur un corpus de parole lue et représentant des situations de la vie quotidienne, car le Score de qualité globale est supérieur à 89. Notre approche reste donc restreinte à ce cadre particulier. Elle n'est pas robuste à des situations acoustiques dites difficiles et à des domaines spécifiques (ce qu'on retrouve dans Propicto-orféo et Propicto-commonvoice). Des pistes de recherche pourraient porter sur l'amélioration de la gestion des termes non traduits et mal traduits. Nous pourrions également tester de nouvelles approches de bout-en-bout et des méthodes pour faire émerger des pictogrammes générés par des systèmes génératifs. Enfin, comparer les sorties entre modèles permettrait d'identifier leurs avantages et inconvénients respectifs.

7 Conclusion

Dans cet article, nous introduisons deux approches pour traduire automatiquement la parole en pictogrammes. Nous présentons des données spécifiquement créées pour cette tâche, couvrant diverses situations acoustiques et divers domaines. Bien que l'approche cascade présente des résultats légèrement supérieurs à l'approche de bout-en-bout sur chaque ensemble de données étudié, nous notons des résultats compétitifs avec cette dernière. Nous n'excluons donc pas cette approche pour des travaux futurs. L'évaluation humaine révèle plusieurs limitations, notamment l'impact significatif des systèmes de reconnaissance vocale sur la traduction, ainsi que la difficulté à traduire certains phénomènes linguistiques tels que les unités polylexicales et les entités nommées. Cette nouvelle tâche Parole-à-Pictos est proposée dans le cadre de la campagne d'évaluation ImageCLEF (Ionescu *et al.*, 2024), intégrée à la conférence CLEF (Conference and Labs of the Evaluation Forum) 2024.

Remerciements

Ce travail a bénéficié d'un financement de l'Agence Nationale de la Recherche, via le projet PROPICTO (ANR-20-CE93-0005). Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2023-AD011013625R1 attribuée par GENCI. Ces travaux ont nécessité l'utilisation de 1 400 heures de GPUs V100, ce qui équivaut à 33 kg de CO_2. Les données utilisées sont libres de droit.

Références

ARDILA R., BRANSON M., DAVIS K., KOHLER M., MEYER J., HENRETTY M., MORAIS R., SAUNDERS L., TYERS F. & WEBER G. (2020). Common voice : A massively-multilingual speech corpus. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 4218–4222, Marseille, France : European Language Resources Association.

ARTETXE M. & SCHWENK H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, **7**, 597–610. DOI : 10.1162/tacl_a_00288.

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.

BARRAULT L., CHUNG Y.-A., MEGLIOLI M. C., DALE D., DONG N., DUPPENTHALER M., DUQUENNE P.-A., ELLIS B., ELSAHAR H., HAAHEIM J. *et al.* (2023). Seamless : Multilingual expressive and streaming speech translation. *arXiv preprint arXiv :2312.05187*.

BAUDE O. & DUGUA C. (2017). Les ESLO, du portrait sonore au paysage digital. *Corpus*. HAL : halshs-01679544.

BENZITOUN C., DEBAISIEUX J.-M. & DEULOFEU H.-J. (2016). Le projet orféo : un corpus d'étude pour le français contemporain. *Corpus*, (15).

BEUKELMAN D. R. & MIRENDA P. (2017). *Communication alternative et améliorée : Aider les enfants et les adultes avec des difficultés de communication*. De Boeck Superieur.

BURCHARDT A. (2013). Multidimensional quality metrics : a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK : Aslib.

CATAIX-NÈGRE E. (2017). *Communiquer autrement : Accompagner les personnes avec des troubles de la parole ou du langage*. De Boeck Superieur.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised cross-lingual representation learning at scale. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440–8451, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.747.

COSTA-JUSSÀ M. R., CROSS J., ÇELEBI O., ELBAYAD M., HEAFIELD K., HEFFERNAN K., KALBASSI E., LAM J., LICHT D., MAILLARD J. *et al.* (2022). No language left behind : Scaling human-centered machine translation. *arXiv preprint arXiv :2207.04672*.

CROIX-ROUGE (2021). Communiquons autrement : Déploiement de la communication alternative améliorée dans les établissements handicap de la croix-rouge française.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DI GANGI M. A., CATTONI R., BENTIVOGLI L., NEGRI M. & TURCHI M. (2019). MuST-C : a Multilingual Speech Translation Corpus. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 2012–2017, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1202.

DUQUENNE P.-A., SCHWENK H. & SAGOT B. (2023). Sonar : sentence-level multimodal and language-agnostic representations. *arXiv e-prints*.

EVAIN S., NGUYEN H., LE H., ZANON BOITO M., MDHAFFAR S., ALISAMIR S., TONG Z., TOMASHENKO N., DINARELLI M., PARCOLLET T., ALLAUZEN A., ESTÈVE Y., LECOUTEUX B., PORTET F., ROSSATO S., RINGEVAL F., SCHWAB D. & BESACIER L. (2021). LeBenchmark : A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. In *INTERSPEECH 2021 : Conference of the International Speech Communication Association*, Brno, Czech Republic. HAL : hal-03317730.

FAN A., BHOSALE S., SCHWENK H., MA Z., EL-KISHKY A., GOYAL S., BAINES M., CELEBI O., WENZEK G., CHAUDHARY V. *et al.* (2021). Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, **22**(107), 1–48.

FENG F., YANG Y., CER D., ARIVAZHAGAN N. & WANG W. (2022). Language-agnostic BERT sentence embedding. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 878–891, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.62.

IONESCU B., MÜLLER H., DRĂGULINESCU A. M., IDRISSI-YAGHIR A., RADZHABOV A., HERRERA A. G. S. D., ANDREI A., STAN A., STORÅS A. M., ABACHA A. B., LECOUTEUX B., STEIN B., MACAIRE C., FRIEDRICH C. M., SCHMIDT C. S., SCHWAB D., ESPERANÇA-RODIER E., IOANNIDIS G., ADAMS G., SCHÄFER H., MANGUINHAS H., COMAN I., SCHÖLER J., KIESEL J., RÜCKERT J., BLOCH L., POTTHAST M., HEINRICH M., YETISGEN M., RIEGLER M. A., SNIDER N., HALVORSEN P., BRÜNGEL R., HICKS S. A., THAMBAWITA V., KOVALEV V., PROKOPCHUK Y. & YIM W.-W. (2024). Advancing multimedia retrieval in medical, social media and content recommendation applications with imageclef 2024. In N. GOHARIAN, N. TONELLOTTO, Y. HE, A. LIPANI, G. MCDONALD, C. MACDONALD & I. OUNIS, Éds., *Advances in Information Retrieval*, p. 44–52, Cham : Springer Nature Switzerland.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2020). BART : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7871–7880, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.703.

LIU Y., GU J., GOYAL N., LI X., EDUNOV S., GHAZVININEJAD M., LEWIS M. & ZETTLEMOYER L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, **8**, 726–742. DOI : 10.1162/tacl_a_00343.

MACAIRE C., DION C., ARRIGO J., LEMAIRE C., ESPERANÇA-RODIER E., LECOUTEUX B. & SCHWAB D. (2024). A multimodal french corpus of aligned speech, text, and pictogram sequences for speech-to-pictogram machine translation. In *LREC*.

MACAIRE C., ESPERANÇA-RODIER E., LECOUTEUX B. & SCHWAB D. (2023). Voice2Picto : un système de traduction automatique de la parole vers des pictogrammes. In C. SERVAN & A. VILNAT, Éds., *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 5 : démonstrations*, p. 10–13, Paris, France : ATALA.

MACAIRE C., ORMAECHEA-GRIJALBA L. & PUPIER A. (2022). Une chaîne de traitements pour la simplification automatique de la parole et sa traduction automatique vers des pictogrammes (simplification and automatic translation of speech into pictograms). In Y. ESTÈVE, T. JIMÉNEZ, T. PARCOLLET & M. ZANON BOITO, Éds., *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 : 24e Rencontres Etudiants Chercheurs en Informatique pour le TAL (RECITAL)*, p. 111–123, Avignon, France : ATALA.

MORRIS A., MAIER V. & GREEN P. (2004). From wer and ril to mer and wil : improved evaluation measures for connected speech recognition.

NORRÉ M., VANDEGHINSTE V., BOUILLON P. & FRANÇOIS T. (2021). Extending a text-to-pictograph system to French and to arasaac. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, p. 1050–1059, Held Online : INCOMA Ltd.

OTT M., EDUNOV S., BAEVSKI A., FAN A., GROSS S., NG N., GRANGIER D. & AULI M. (2019). fairseq : A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019 : Demonstrations*.

OTT M., EDUNOV S., GRANGIER D. & AULI M. (2018). Scaling neural machine translation. In O. BOJAR, R. CHATTERJEE, C. FEDERMANN, M. FISHEL, Y. GRAHAM, B. HADDOW, M. HUCK, A. J. YEPES, P. KOEHN, C. MONZ, M. NEGRI, A. NÉVÉOL, M. NEVES, M. POST, L. SPECIA, M. TURCHI & K. VERSPOOR, Éds., *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 1–9, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-6301.

PEREIRA J., NOGUEIRA R., ZANCHETTIN C. & FIDALGO R. (2023). Predictive authoring for brazilian portuguese augmentative and alternative communication. *arXiv preprint arXiv :2308.09497*.

PEREIRA J. A., DE MEDEIROS S., ZANCHETTIN C. & FIDALGO R. D. N. (2022a). Pictogram prediction in alternative communication boards : a mapping study. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, p. 705–717 : SBC.

PEREIRA J. A., MACÊDO D., ZANCHETTIN C., DE OLIVEIRA A. L. I. & DO NASCIMENTO FIDALGO R. (2022b). Pictobert : Transformers for next pictogram prediction. *Expert Systems with Applications*, **202**, 117231.

POST M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 186–191, Belgium, Brussels : Association for Computational Linguistics.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, p. 28492–28518 : PMLR.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, **21**(140), 1–67.

RAVANELLI M., PARCOLLET T., PLANTINGA P., ROUHE A., CORNELL S., LUGOSCH L., SUBAKAN C., DAWALATABAD N., HEBA A., ZHONG J., CHOU J.-C., YEH S.-L., FU S.-W., LIAO C.-F., RASTORGUEVA E., GRONDIN F., ARIS W., NA H., GAO Y., MORI R. D. & BENGIO Y. (2021). SpeechBrain : A general-purpose speech toolkit. arXiv :2106.04624.

SEVENS L. (2018). *Words divide, pictographs unite : Pictograph communication technologies for people with an intellectual disability*. Netherlands Graduate School of Linguistics.

SEVENS L., VANDEGHINSTE V., SCHUURMAN I. & VAN EYNDE F. (2015). Extending a Dutch text-to-pictograph converter to English and Spanish. In J. ALEXANDERSSON, E. ALTINSOY, H. CHRISTENSEN, P. LJUNGLÖF, F. PORTET & F. RUDZICZ, Éds., *Proceedings of SLPAT 2015 : 6th Workshop on Speech and Language Processing for Assistive Technologies*, p. 110–117, Dresden, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/W15-5119.

VANDEGHINSTE V., SEVENS I. S. L. & VAN EYNDE F. (2017). Translating text into pictographs. *Natural Language Engineering*, **23**(2), 217–244.

VASCHALDE C., TRIAL P., ESPERANÇA-RODIER E., SCHWAB D. & LECOUTEUX B. (2018). Automatic pictogram generation from speech to help the implementation of a mediated communication. In *Conference on Barrier-free Communication*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30 : Curran Associates, Inc.

WANG C., TANG Y., MA X., WU A., OKHONKO D. & PINO J. (2020). Fairseq S2T : Fast speech-to-text modeling with fairseq. In D. WONG & D. KIELA, Éds., *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing : System Demonstrations*, p. 33–39, Suzhou, China : Association for Computational Linguistics.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., LE SCAO T., GUGGER S., DRAME M., LHOEST Q. & RUSH A. (2020). Transformers : State-of-the-art natural language processing. In Q. LIU & D. SCHLANGEN, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-demos.6.

WOODARD J. & NELSON J. (1982). An information theoretic measure of speech recognition performance.

YE R., WANG M. & LI L. (2021). End-to-End Speech Translation via Cross-Modal Progressive Training. In *Proc. Interspeech 2021*, p. 2267–2271. DOI : 10.21437/Interspeech.2021-1065.

YE R., WANG M. & LI L. (2022). Cross-modal contrastive learning for speech translation. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 5099–5113, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.376.