A Multimodal French Corpus of Aligned Speech, Text, and Pictogram Sequences for Speech-to-Pictogram Machine Translation

Cécile Macaire, Chloé Dion, Jordan Arrigo, Claire Lemaire, Emmanuelle Esperança-Rodier, Benjamin Lecouteux, Didier Schwab

A Multimodal French Corpus of Aligned Speech, Text, and Pictogram Sequences for Speech-to-Pictogram Machine Translation

Cécile Macaire[1], **Chloé Dion**[1], **Jordan Arrigo**[1], **Claire Lemaire**[1,2], **Emmanuelle Esperança-Rodier**[1], **Benjamin Lecouteux**[1], **Didier Schwab**[1]
[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France,
[2]LAIRDIL, IUT, Univ. Paul Sabatier, 115 B rte de Narbonne, 31077 Toulouse, France
[1,2]first.last@univ-grenoble-alpes.fr

Abstract

The automatic translation of spoken language into pictogram units can facilitate communication involving individuals with language impairments. However, there is no established translation formalism or publicly available datasets for training end-to-end speech translation systems. This paper introduces the first aligned speech, text, and pictogram translation dataset ever created in any language. We provide a French dataset that contains 230 hours of speech resources. We create a rule-based pictogram grammar with a restricted vocabulary and include a discussion of the strategic decisions involved. It takes advantage of an in-depth linguistic study of resources taken from the ARASAAC website. We validate these rules through multiple post-editing phases by expert annotators. The constructed dataset is then used to experiment with a Speech-to-Pictogram cascade model, which employs state-of-the-art Automatic Speech Recognition models. The dataset is freely available under a non-commercial licence. This marks a starting point to conduct research into the automatic translation of speech into pictogram units.

Keywords: Pictograms, Speech, Machine Translation

1. Introduction

Augmentative and Alternative Communication (AAC) encompasses tools and strategies to facilitate communication when the conventional language abilities of an individual are impaired or absent (Cataix-Nègre, 2017). Language impairment (oral production and comprehension, listening, reading, and writing) can arise from various sources, including genetic diseases, autism spectrum disorders, or intellectual disability, among others. AAC offers a range of tools, such as communication boards, signs, and computer devices, to help individuals effectively convey their messages in everyday situations (Romski and Sevcik, 2005).

One of the features of AAC involves the representation of natural language into pictograms, a graphic representation associated with a concept (object, person, action, etc.) (Pereira et al., 2022b). The use of pictograms as a communication aid has demonstrated its effectiveness in visualizing syntax, manipulating words, and enhancing language accessibility (Cataix-Nègre, 2017). Additionally, a study conducted by the French Red Cross identified several positive outcomes associated with AAC (Croix-Rouge, 2021). These include reduced stress, increased autonomy and health, as well as heightened calmness and enjoyment in daily life.

However, there is a need to offer substantial support for these technologies, both from the standpoint of the AAC users, and from the perspective of their families and caregivers (Beukelman and Mirenda, 2013). There are several environmental hurdles to overcome in the implementation of AAC technologies, including a lack of visibility and availability, particularly in light of the considerable time and effort required for tool calibration and adaptation, which can span several months. In addition, caregivers are not accustomed to using pictograms (Moorcroft et al., 2019). We argue that providing Speech-to-Pictogram (S2P) translation systems could address these challenges by enhancing communication for AAC users.

Previous work has focused on the text-to-pictogram task (Sevens et al., 2015; Vandeghinste et al., 2017; Sevens, 2018; Norré et al., 2021). However, when the input is spoken language, it becomes imperative to consider the intricacies inherent in orality. Spoken language tends to contain certain disfluencies (vocal hesitations, discursive markers, self-corrections, false starts, etc.), which are integral to spontaneous speech but do not carry inherent meaning. Therefore, they cannot be translated into pictograms. This NLP task which aims to directly transcribe spoken utterances to a sequence of corresponding pictograms faces several challenges:

- the non-existence of parallel speech-pictogram corpora, which limits the use of state-of-the-art end-to-end machine learning architectures;
- the difficulty of translating into pictograms when standardized rules have not been established (how to handle negation, proper nouns, and tense markers, among others);
- the absence of terms translated into pictograms and the ambiguity in labelling terms with pictograms, undermining the quality of the translation.

In this paper, we describe our efforts to collect and create a large corpus of aligned speech to pictogram units. We present the main features and provide an in-depth presentation of its development process, from crawling and cleaning, to construction and evaluation. We introduce baseline NLP systems for the Speech-to-Pictogram (S2P) task based on a cascade architecture. We summarize our **contributions** below[1]:

- The definition of a formalization to arrange pictograms, based on a linguistic study, and which will be then considered as a translation reference,
- The construction and release of a large French dataset for the S2P task, Propicto-orféo [2],
- Baseline systems that are readily available to use, followed by an initial evaluation of their performances.

We discuss the relevant research related to this work in Section 2. Section 3 then outlines the methodology employed to create and validate the dataset, along with its key statistics. In Section 4, we detail the experimental setup used for baseline models to give an initial evaluation of the created dataset on the Speech-to-Pictogram task. The results are presented in Section 5, followed by a short summary of the research and future work in Section 6.

2. Related Work

The process of translating French speech into a sequence of pictogram units is investigated in Vaschalde et al. (2018). In their work, they adapt the Text2Picto system proposed by Vandeghinste et al. (2017) for text-to-pictogram translation to accommodate speech input. The system includes four modules: an Automatic Speech Recognition (ASR) system, a simplification module, a disambiguation model and a display module to output the correct ARASAAC pictogram. To assess the system's performance, two datasets are employed. The first dataset consists of fifteen children's stories, each manually translated, while the second dataset features 20 sentences extracted from the ESLO corpus (Baude and Dugua, 2008), for a total of 2,000 pictograms. This represents a low-resource scenario.

To our knowledge, prior research has focused on the translation dimension from text to pictograms rather than from speech to pictograms. Sevens et al. (2015) introduced a Text-to-Picto system for Dutch, later adapted to handle English and Spanish (Sevens, 2018) within the Able to Include project [3]. The system first performs a linguistic analysis, which includes tokenization, part-of-speech tagging, lemmatization, Named Entity Recognition (NER) and multi-word expression processing. A Word Sense Disambiguation (WSD) module completes the pipeline by retrieving a specific sense for each term. Norré et al. (2021) further extend this work to include French and ARASAAC pictograms. The evaluation draws from three distinct corpora: the email corpus, consisting of 130 sentences manually translated into Sclera and Beta pictograms by Sevens (2018); the book corpus, comprising 254 sentences in ARASAAC pictograms as detailed in Vaschalde et al. (2018); and the medical corpus, containing 260 sentences featuring medical terms (questions from doctors to patients and patient instructions).

It is worth mentioning a few works on predicting pictograms for AAC systems (Pereira et al., 2022a, 2023). The objective is to predict the correct pictogram given the context. Pereira et al. (2023) adapt BERT for pictogram prediction specifically for Brazilian Portuguese. To overcome the shortage of resources for fine-tuning such architectures, they compile a dataset of AAC-like sentences from AAC practitioners, representing a total of 667 sentences. They employ GPT-3 (Brown et al., 2020) for data augmentation to generate synthetic sentences.

Pictogram translation remains unexplored, with a community still confronted with scattered and non-standardized resources. The availability of large training corpora is crucial to the development of end-to-end architectures.

3. Resource Creation

This section explains the process of constructing the dataset of aligned speech and pictogram sequence, from the data collection (Section 3.1) to its construction (Section 3.2) and evaluation (Section 3.3). We give an overview of the statistics in Section 3.4.

3.1. Dataset Collection

We explore an approach which generates a sequence of pictograms from oral transcriptions, using rule-based grammar.

We begin by collecting annotated French texts in pictograms, sourced from Norré et al. (2021). They are used to carry out a linguistic study of pictogram translation patterns to define a formalism. This dataset comprises three distinct corpora, containing sentences taken from children's stories, medical contexts, and emails. For a representative range

[1] Code released at https://github.com/macairececile/picto_grammar/

[2] Freely available here: https://www.ortolang.fr/market/corpora/propicto

[3] https://able-to-include.ccl.kuleuven.be/index.html

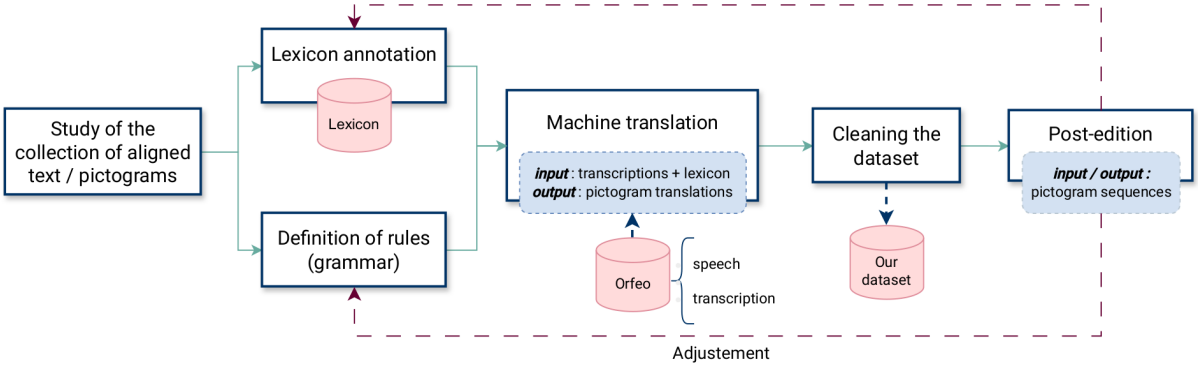

Figure 1: Dataset construction pipeline.

of translations and domains, we integrate an additional set of 304 sentences. These are extracted from freely available PDF files on the ARASAAC website[4] with content from children's stories, poems, songs, and everyday life situations. All the resources have been carefully compiled and annotated by speech and language therapists. Table 1 presents quantitative statistics regarding this collection. It consists of 897 sentences translated into pictogram units, with over 5,800 pictograms and 7,800 words.

Resource	# Sentences	# Words	# Pictos
Medical corpus	260	1,757	1,490
Book corpus	253	2,544	2,207
Email corpus	80	907	750
Arasaac corpus	304	2,620	1,392
Resource	**# Sentences**	**# Words**	**Duration (h)**
Orféo	291,272	2,940,831	234

Table 1: Data collection.

Next, we leverage French speech data provided by the corpus C.E.F.C (Benzitoun et al., 2016) from the Orféo project (Benzitoun and Debaisieux, 2020) with a pictogram translation. This freely accessible resource comprises 12 sub-corpora with more than 2,500 different speakers, from a variety of speech situations: conversation, meetings, child's speech, etc. Orféo is a corpus of spontaneous speech, with syntactic-level annotations, providing information to map text to pictograms. This collection is valuable, by the diversity of speech situations, topics, and speaker profiles. We can explore a range of acoustic, domain-specific, and translation scenarios, enhancing the versatility and applicability of our approach.

It is worth noting that all the resources in our study employ ARASAAC pictograms. This knowledge database is freely available under a Creative Commons license (BY-NC-SA), with a collection of 25,000 pictograms. With its multilingual support, high coverage of terms, and ongoing expansion of new pictograms, ARASAAC makes it one of the most widely used resources in the AAC community[5].

3.2. Dataset Construction

The pipeline to construct aligned French speech, text, and pictogram sequences is presented in Figure 1. Our methodology can be broken down into 3 key points:

1. The creation of a lexicon, whose goal is to map a term to a unique pictogram identifier. We define a *term* as a word, a multi-word expression or a full sentence.

2. The study of the collected texts translated into pictogram units (see Section 3.1) to extract the characteristics and formalize a grammar.

3. The implementation of a machine translation rule-based system which is based on the formalism defined in the previous step.

3.2.1. Lexicon Creation

The ARASAAC resource contains approximately 25,000 pictograms, each of which is annotated with a unique identifier and a set of keywords. A single keyword can be linked to multiple pictograms, as shown in Figure 2, where the term '*cheval*' ('horse') is linked to 3 pictograms. Two refer to the animal, whereas one corresponds to the knight in chess.

On the contrary, terms do not have associated pictograms, for example, *'rencontrer'* ('to meet'). It is therefore necessary to develop a restricted lexicon that takes these challenges into account, ensuring improved word coverage and reliable translations.

[4] https://arasaac.org/

[5] https://aulaabierta.arasaac.org/en/visual-arasaacs-map

Figure 2: ARASAAC pictograms linked to the term *'cheval'* ('horse').

To create a set of locutions linked to pictograms, we investigated two main resources. The first one is texts that are annotated with a sequence of pictograms, presented in Section 3.1. From these texts, we extract terms with a corresponding pictogram translation. The second resource is a set of speech corpora from which we obtain terms and formulations commonly used when speaking (hesitations, fillers, etc.). The spoken utterances and transcriptions are taken from three distinct corpora: REPERE (Giraudel et al., 2012) (TV shows concerning news and debates), ETAPE (Gravier et al., 2012) (TV and radio broadcasts), and ESTER1 (Gravier et al., 2004) (radio station recordings). Additionally, we gather a set of sentences from the TCOF corpus (André and Canut, 2010). It contains recordings of adult-child (children up to age 7) and adult-adult interactions. We also use a text corpus containing questions asked on medical forums and exchanges occurring during medical consultations. The created lexicon covers commonly used terms in the French language, in everyday life interactions, as well as in specialized fields (medicine, news).

Locution	Pictogram identifier
astronome # masculin 'astronomer'	32085
vous #neutre 'you'	7307
je ne suis pas d'accord 'I do not agree'	37182

Table 2: Sample of the lexicon.

Table 2 gives a sample of the lexicon. The first column refers to the locution, which can be all types of terms, sentences, and multi-word expressions. Each locution is linked to a pictogram identifier. For some, we incorporate tags (indicated by #) which categorize and differentiate possible homonyms into 4 groups: (1) grammatical category (#adjective, #interrogative, etc.), (2) gender (#feminine, #masculine, #neutral), (3) #animal, and (4) #person/#object. The use of one pictogram will obviously depend on the context of use. The lexicon contains 3,785 terms, which is subsequently merged with the lexicon defined in Norré et al. (2021), resulting in a final resource containing 26,655 annotated terms.

3.2.2. Definition of Rules

We define the set of rules to associate each French spoken language transcription into a sequence of corresponding pictograms. Each resulting pictogram is linked to a unique term (also referred to as the 'picto token'). We then implement them in an automatic manner (see Section 3.2.3).

To begin with, we extract the characteristics of a translation into pictograms by studying the collection of texts presented in Section 3.1. Through this process, we identified three levels of translation, relying on the grammatical categories of translated words. Level 1 translates only lexical words, i.e. common nouns, specific proper nouns, and verbs. Adjectives and adverbs are not systematically translated. Level 2 takes into account the pronouns in addition to the lexical words, making it possible to provide information about the gender (feminine, masculine, neutral) and number (singular/plural). Finally, in level 3, there is a will to translate every term in the text.

We construct the rules according to a breakdown of linguistic phenomena into 4 parts:

- The syntactic units, which concern the translation of terms, but also the different types of sentences (interrogative, exclamatory, declarative and imperative), and multi-word expressions.
- The construction of the verbal nucleus, which aims to define the rules concerning conjugation, the passive voice and negation.
- The grammatical categories, i.e. person, gender, number, mode, and tense.
- The derivational morphology, particularly translation of affixes.

The rules, 10 in total, described below, build a level 3 translation. They can be easily adjusted to translate to level 1 or 2. For each, we give a concrete example with the transcription, the token of each pictogram chosen by the grammar, and the visual sequence of pictograms. The element to which the rule refers is highlighted in bold.

- **Tense marker** — a pictogram referring to the past or the future is added to the translation before the subject of the verbal group.

text		Il	**mangea**	tard
token	**passé**	il	manger	tard
picto				
transl	'He ate late.'			

- **Imperative** — an imperative marker is placed at the end of an imperative sentence with the pictogram '!'.

text token	Partez partir	maintenant maintenant	!
picto			
transl	'Leave now.'		

- **Pronominal verb** — the pictogram linked to the pronominal verb (verb accompanied by a reflexive pronoun) is retrieved, if it exists in the lexicon. If not, we display the pictogram associated to the verb only.

text token	Je je	**m'habille** s'habiller	rapidement rapidement
picto			
transl	'I get dressed quickly.'		

- **Singular/Plural** — the pictogram associated with the plural term is shown if it exists in the lexicon, otherwise the pictogram associated with the lemmatized form is retrieved.

text token	Vous vous	achetez acheter	des des	**fruits** fruits
picto				
transl	'You buy fruit.'			

- **Number** — numbers are translated into a unique pictogram, if it exists in the lexicon.

text token	J'amène je amener	**soixante dix** 70	beignets beignets
picto			
transl	'I'm bringing sixty-six doughnuts.'		

- **Negation** — a negation marker is put after the verbal group to which it relates with a specific negation pictogram. Other terms referring to negation are not translated.

text token	On nous	**ne** Ø	peut pouvoir	**pas** dialoguer	dialoguer non
picto					
transl	'We can't talk to each other.'				

- **Prefix** — the term of the form prefix+word is split into two pictograms if the prefix is one of the X identified: a pictogram related to the prefix, and a pictogram linked to the term.

text token	Ce le	parking parking	est être	**inaccessible** in	accessible
picto					
transl	'This car park is inaccessible.'				

- **Multi-word expression** — the grammar looks at up to 8 consecutive words to detect multi-word expressions. The pictogram is printed if it exists in the lexicon.

text token	Je je	**me brosse les dents** se_brosser_les_dents
picto		
transl	'I brush my teeth.'	

- **Untranslated expressions** — the terms belonging to the list of well known French disfluencies are deleted.

text token	**Hum** Ø	**euh** Ø	d'accord d'accord
picto			
transl	'Um um all right.'		

3.2.3. Machine Translation

The pipeline to automatically generate a translation in pictograms is described in Figure 3. It first employs a spaCy model[6] to perform tokenization, lemmatization, and dependency parsing. This information is useful to retrieve specific linguistic features (gender, singular, plural, negation, tense of the verbs) required to apply these rules.

[6] https://spacy.io/

audio	cefc-tcof-Mili_89-285					
sentence	eh ben c'était un le tremblement de terre					
pictos_tokens	passé (9839)	celui-là (7095)	être (36480)	un (2627)	le (8476)	séisme (4755)
arasaac_pictos						
transl	'eh well it was an the earthquake'					

Table 3: An example taken from the subcorpus Tcof, with the spoken utterance name, the aligned transcription and the sequence of pictograms generated by the grammar.

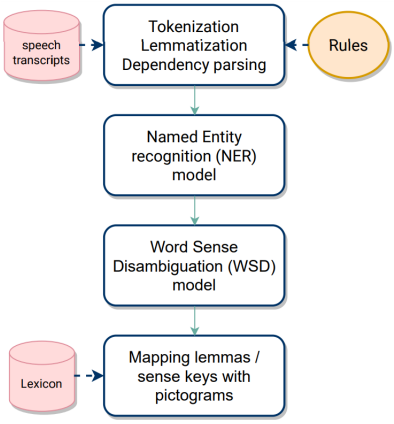

Figure 3: Machine translation pipeline.

The lexicon has a limited amount of annotated named entities. To translate those that are not, one solution is to apply a Named Entity Recognition (NER) model that recognizes them, and associates a pictogram by default according to the category. Therefore, we apply a French NER model entitled *camembert-ner*[7] fine-tuned from camemBERT (Martin et al., 2020) on the wikiner-fr dataset (Nothman et al., 2013). This model obtains a F-score of 89%. It tags the input text into 5 categories, from which we only keep 3 on our grammar: PER (person's name), ORG (organization), and LOC (location). For each, we assign a specific pictogram.

To handle terms that do not have any related pictogram, we use a neural classifier for the task of Word Sense Disambiguation (WSD) proposed in Vial et al. (2019). The goal is to retrieve the correct sense given the context. Here, each pictogram in ARASAAC is linked to a WordNet 3.1 sense key. The architecture takes contextualized word vectors extracted from a pre-trained language model. These vectors are fed into a series of 6 Transformer encoder layers (Vaswani et al., 2017). A softmax layer predicts the WordNet sense key (Miller, 1995) of each input word. The best model obtains an F-score of 56.95% on the French part of SemEval 2013 (Vial, 2019). The architecture is trained on two datasets: SemCor (Miller et al., 1993) and the WordNet Gloss Corpus (WNGT) with 155,125 sentences as the training set, and 4,000 as the validation set (Vial, 2019). Finally, we construct the sequences of pictograms thanks to the lexicon and the information annotated by the grammar.

An example of a translation is presented in Table 3. For each speech utterance, the aligned transcription is given, as well as the sequence of pictograms. We also give the tokens linked to each of them, this information is taken from the lexicon. The arasaac_pictos line is only shown here to give a visual example to the reader.

3.3. Dataset Cleaning and Validation

For both corpora, we exclude speech utterances that do not contain any words. We also remove audio segments that do not have a provided pictogram sequence. This includes speech composed of disfluencies such as fillers, hesitations, and verbal tics. For instance, the speech utterance 'cefc-clapi-montage_meuble-23' with the content 'hm voilà' (meaning 'hm here it is') falls into this category. In total, these instances account for less than 0.004% of the audio clips.

In an attempt to assess the performance of the grammar, we perform a post-edition step. This refers to the task of human review of the translation performed by an automatic system (Robert, 2010). Commonly used in machine translation (Koponen, 2016), we adapt the process to pictograms. The aim of this phase is to verify the accuracy of the generated pictogram sequence within a set of speech utterances and to update the formalism based on observations. In an online platform that we developed, two expert annotators from the project were given the transcription and the translation in pictograms. Specific guidelines were provided for post-editing. The first one is to make any edits deemed necessary to match the defined formalism.

[7] https://huggingface.co/Jean-Baptiste/camembert-ner

	# Utterances	Duration (in hours)	# Speakers	# Words	# Unique words	# Pictos	# Unique pictos	Speech type
cfpb	4,953	3,71	8 (3 / 5 / -)	47,463	4,592	39,042	1,755	Spontaneous
cfpp	41,582	32,78	74 (29 / 43 / 2)	401,832	15,604	342,699	3,815	Spontaneous
clapi	20,794	10,26	125 (24 / 35 / 66)	141,090	9,265	119,751	2,908	Spontaneous
coralrom	18,422	17,56	158 (82 / 75 / 1)	207,980	15,127	175,721	3,534	Spontaneous
crfp	29,965	28,81	19 (14 / 4 / 1)	343,248	17,950	290,589	4,299	Spontaneous
fleuron	2,875	2,26	8 (2 / 6 / -)	25,915	2,330	21,999	1,020	Spontaneous
frenchoralnarrative	12,745	11,84	21 (9 / 10 / 2)	129,819	10,610	119,128	2,947	Read
ofrom	21,566	19,66	99 (62 / 137 / -)	246,650	13,681	205,994	3,787	Spontaneous
reunions	16,566	13,78	15 (9 / 5 / 1)	168,763	9,055	140,045	2,224	Spontaneous
tcof	27,099	21,33	13 (4 / 8 / 1)	291,371	14,962	246,122	4,048	Spontaneous
tufs	56,896	39,35	99 (37 / 61 / 1)	547,117	19,250	469,366	4,516	Spontaneous
valibel	36,573	32,45	263 (163 / 94 / 6)	386,743	18,805	330,715	3,757	Spontaneous
all	**290,036**	**233,79**	**1002 (438 / 483 / 81)**	**2,937,991**	**57,657**	**2,501,171**	**6,503**	**Mixed**

Table 4: Statistics of the created dataset of aligned speech/text/pictograms. The number of speakers is broken down by gender (male / female / unknown).

The second guideline is to correct any pictogram representing a term that does not match the defined lexicon. Annotators had the choice to remove, add, or change the order of each pictogram. A set of 100 sentences per subcorpus constituting the created dataset was post-edited, representing a total of 1,200 utterances.

		$Expert_1$		
		Yes	No	
$Expert_2$	Yes	749	159	908
	No	88	204	292
	Total	837	363	1200

Table 5: Confusion matrix between the two annotators.

Table 5 presents the confusion matrix of agreement between the two annotators. The “yes” refers to the annotator accepting the translation without any edits, whereas “no” means at least one edit was performed on the translation. For 83% of sentences, at least one annotator accepts the provided sequence of pictograms without any edits. This first measure shows that the majority of machine-translated sentences are reliable. The rules address most translation scenarios, and the created lexicon proposes accurate and coherent pictograms.

For a more precise study, we calculate the Translation Edit Rate (TER) (Snover et al., 2006; Post, 2018), which quantifies the number of edit operations needed for a hypothesis to align with a reference translation. Based on the average pictogram translation length of 8.5 pictograms in the corpus, and with a TER score of 6.36%, less than one term is edited (precisely 0.54). In Table 6, we present statistics of the number of edits carried out by annotators according to their types. Deletions emerge as the most frequent type of edit for both experts, indicating a tendency in the formalism to translate a term that should be represented by a single pictogram (such as multi-word expressions or named entities) into multiple ones. Substitutions come next, often attributed to mistranslations of terms, while insertions, stemming from a pictogram omission, occur with less frequency.

	# insertion	# deletion	# substitution
$Expert\_1$	104	298	144
$Expert\_2$	52	259	81

Table 6: Number of edit per type and per annotator.

We also compute the inter-annotator agreement, a good indicator for measuring how close the decisions between annotators are. The annotators made the same decision on 953 sentences, for a proportionate agreement of 79.42%. Cohen’s Kappa is of 0.48, which corresponds to a moderate agreement.

These results can be explained by the difficulty of the choices the annotators had to make, when a term had never been translated into a pictogram (not the same pictogram between annotators), or when an annotator decided to remove or keep a specific pictogram, while the other did not deem it necessary to do so.

3.4. Dataset Statistics

Statistics of the final dataset, which makes it the first aligned speech, text, and pictograms one for the French language, are printed in Table 4. It contains over 233 hours of speech, from 1,002 unique speakers, consisting of 290,036 utterances. The dataset gathers, in majority, spontaneous speech, except one subcorpus, the ‘frenchoralnarrative’ with read speech. The dataset Propicto-orféo is freely available under a non-commercial licence.

Corpus →	cfpb	cfpp	clapi	coralrom	crfp	fleuron	narrative	ofrom	reunions	tcof	tufs	valibel	all
WER (%)	40.8	47.1	70.5	34.4	37.1	43.8	16.6	33.6	58.5	47.1	47.1	37.6	**43.0**
BLEU	55.9	55.4	30.6	65.3	62.0	58.3	80.2	61.7	38.9	49.4	51.5	57.4	57.7
METEOR	60.1	60.4	31.4	69.2	64.1	55.1	83.3	71.3	45.5	56.6	59.7	64.1	60.9
PER (%)	36.4	42.0	66.6	28.7	32.1	40.1	12.9	28.3	54.1	43.3	40.8	34.1	38.1

Table 7: WER (%) score on our test set by *whisper-large* ASR model. We give the BLEU, METEOR and PER (%) between the sequence of reference pictogram tokens and the one predicted from the ASR hypothesis on the test set.

4. Experiments

We conduct experiments on the Speech-to-Pictogram task to establish the usability of our dataset. Our objective is to assess the performance of an Automatic Speech Recognition (ASR) system on a test set, and understand its impact on the pictogram translation.

ASR is the task of retrieving what was said in a spoken utterance into text. In this work, we try several architectures. Wav2Vec2.0 is a Self-Supervised learning (SSL) framework (Watanabe et al., 2017; Baevski et al., 2020) which learns powerful speech representations from a huge collection of unlabeled speech (during pretraining) followed by a fine-tuning step on transcribed speech for a downstream task. More recently, Whisper (Radford et al., 2023) was introduced as a multilingual ASR system with the capability to produce competitive results on robustness (accents, background noises) and accuracy, without any fine-tuning. Based on an encoder-decoder Transformer architecture (Vaswani et al., 2017), the model was trained on 680,000 hours of multilingual and multitask supervised online data.

The predictions given by the best ASR model are passed through the grammar to compare the reference pictogram translation (i.e. the sequence of terms linked to each pictogram) with the predicted one.

For all experiments, we employ a test set from our corpus, comprising 10% of the data from each subcorpus, totaling 23 hours and 29,214 utterances.

5. Results and Discussion

5.1. Automatic Evaluation

We compare a French Wav2Vec2.0 model with several Whisper models. We present the Word Error Rate (WER) per subcorpus from the best model in Table 7. In our test set, *whisper-large*[8] gives the best performance with a WER of 43%. We hypothesize that due to Whisper's training on an extensive quantity of data, it grasps a greater amount of linguistic information. A disparity between corpora is noticeable, particularly when comparing the scores between clapi and frenchoralnarrative. This observation is not surprising, as the first contains very noisy speech, with overlap between speakers, while the other is read speech in an optimal acoustic environment. We do not exclude the possibility that performances could be improved by using other ASR models. However, these results give an initial estimate of the performance we can expect from a challenging corpus, in a spontaneous environment.

In the same table, we show the BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and Picto Error Rate (PER) (%) scores by comparing the reference grammar translation with the one generated by the grammar from the predicted ASR transcriptions. We recall that the reference grammar translation is the final sequence of pictogram terms generated by the grammar and validated through post-edit sessions. We define the PER as the number of substitutions (S), deletions (D), and insertions (I) to match the predicted sequence of pictogram tokens with the reference one. The BLEU score of 57.7 on the test set emphasizes less degradation of the translation into pictograms from the ASR output. The lowest BLEU score is encountered when the WER score is the highest (e.g. clapi), also true in the other direction.

WER (%)	BLEU	METEOR	PER (%)	Data proportion (%)
0 – 9.99	**92.4**	**96.3**	**3.8**	**12.2**
10 – 19.99	81.4	90.3	11.9	12.2
20 – 29.99	71.0	83.2	20.1	15.2
30 – 39.99	61.5	75.6	28.3	9.8
40 – 49.99	52.0	68.8	36.1	8.4
> 50	21.7	28.6	79.9	42.0

Table 8: BLEU, METEOR, and PER (%) between the sequence of reference picto tokens and the one predicted from the ASR hypothesis on the test set, based on a WER scale.

We present a final study to assess how ASR performance affects the pictogram translation. As shown in Table 8, a WER ranging from 0 and 9.99% appears to have minimal impact on the translation quality. The observation holds true up to a WER of 30%. Beyond this threshold, a noticeable decline in results is seen, with a significant decrease in the BLEU score when the WER is over 50%.

[8] https://huggingface.co/openai/whisper-large

However, despite a notably poor WER score, the pictogram translation is relatively unaffected, as evidenced by a BLEU score of 52 when the WER falls between 40 and 49.9%. Specifically, a 52 BLEU score indicates a translation in which certain pictograms have been altered, making it understandable to the user but potentially leading to misinterpretation.

5.2. Human Evaluation

We conduct a human evaluation to provide a precise analysis of the types of errors generated by the ASR system and its impact on pictogram translation. We adapt an analytical framework, MQM (Burchardt, 2013), which offers guidelines and procedures for measuring translation quality [9]. This framework determines whether the proposed translation meets the specifications agreed upon by stakeholders. Each expert annotator assigns a specific type and severity level to each identified error in the text (source and/or target).

In this study, two experts annotated 100 randomly chosen sentences per sub-corpus from the Orféo-picto test. This results in 1,200 sentences, corresponding to a WER score of 40.1%. The reference grammar translation was provided alongside the pictogram translation generated by the ASR hypothesis. The annotators' goal was to compare the two and annotate each encountered error from a list of 12 error types (addition, omission, unintelligible, etc.) and 4 severity levels (neutral, minor, major, critical).

Severity → **Error types ↓**	**Neutral**	**Minor**	**Major**	**Critical**	**Total**
Accuracy					
Mistranslation	156	163	317	334	970
Addition	163	150	72	34	419
Omission	519	330	296	155	1300
Over-translation	0	29	6	0	35
Under-translation	0	1	0	0	1
Fluency					
Word-order	27	13	3	0	40
Offensive	0	0	1	0	1
Unintelligible	0	0	12	287	299

Table 9: Number of errors per category and per severity level annotated from both experts according to MQM framework.

Table 9 displays the number of annotated errors categorized by type and severity level from both experts. Critical errors include mistranslation, addition, omission, and unintelligible errors. The ASR hypothesis may mistranslate or omit parts of the speech, resulting in the absence, inaccuracy, or addition of pictograms. A critical mistranslation typically involves terms with significant semantic importance, such as verbs or nouns. Errors categorized as unintelligible are always considered major or critical, indicating an ASR hypothesis with a Word Error Rate (WER) score exceeding 60%. Addition or omission errors generally have a lesser impact on translation, often falling into the neutral or minor severity levels. These errors typically involve the addition or omission of less semantically significant elements, such as demonstrative pronouns, temporal markers, adverbs, or punctuation.

[9] https://themqm.org/

6. Conclusion and Future Work

In this paper, we present a dataset for the task of Speech-to-Pictogram translation to enhance communication for AAC users. We propose a rule-based grammar and a restricted lexicon, taking as input oral transcriptions for the construction of the first large aligned speech/text/pictograms corpus. Post-editing has shown that our methodology produces reliable translations. The experiments which employ state-of-art ASR models combined with the grammar exhibit that the translation quality is not greatly affected by the performance of the ASR, which is not true for other tasks. A more in-depth study is required to confirm our observations.

We aim to extend the dataset to other speech corpora, and develop multilingual and multitask speech-to-text architectures on our dataset. Finally, we want to incorporate a qualitative evaluation with our real-world users of this technology to offer fine-grained and adjusted translations. In particular, we could consider using cutting-edge technologies, such as eye-trackers or EEG headsets to assess their ability to understand the sequence in pictograms.

Acknowledgements

This project was funded by the Agence National de la Recherche (ANR) through the project PROPICTO (ANR-20-CE93-0005). We thank Aidan Mannion and Ange Richard for their valuable feedbacks.

7. Bibliographical References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic*

and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

David R Beukelman and Pat Mirenda. 2013. *Augmentative and alternative communication: Supporting children and adults with complex communication needs*. Paul H. Brookes Pub.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

Elisabeth Cataix-Nègre. 2017. *Communiquer autrement: Accompagner les personnes avec des troubles de la parole ou du langage*. Apprendre et réapprendre. De Boeck Supérieur.

Croix-Rouge. 2021. Communication alternative améliorée (caa) : la croix-rouge française dévoile sa première étude d'impact social !

Maarit Koponen. 2016. Is machine translation post-editing worth the effort? a survey of research into post-editing and effort. *The Journal of Specialised Translation*, 25(2).

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

A Moorcroft, N Scarinci, and C Meyer. 2019. A systematic review of the barriers and facilitators to the provision and use of low-tech and unaided aac systems for people with complex communication needs and their families. *Disability and Rehabilitation: Assistive Technology*, 14(7):710–731.

Magali Norré, Vincent Vandeghinste, Pierrette Bouillon, and Thomas François. 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)*, pages 1050–1059, Held Online. INCOMA Ltd.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175. Artificial Intelligence, Wikipedia and Semi-Structured Resources.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jayr Pereira, Rodrigo Nogueira, Cleber Zanchettin, and Robson Fidalgo. 2023. Predictive authoring for brazilian portuguese augmentative and alternative communication. *arXiv preprint arXiv:2308.09497*.

Jayr A Pereira, Sheyla de Medeiros, Cleber Zanchettin, and Robson do N Fidalgo. 2022a. Pictogram prediction in alternative communication boards: a mapping study. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 705–717. SBC.

Jayr Alencar Pereira, David Macêdo, Cleber Zanchettin, Adriano Lorena Inácio de Oliveira, and Robson do Nascimento Fidalgo. 2022b. Pictobert: Transformers for next pictogram prediction. *Expert Systems with Applications*, 202:117231.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Anne-Marie Robert. 2010. La post-édition: l'avenir incontournable du traducteur? *Traduire. Revue française de la traduction*, (222):137–144.

MaryAnn Romski and Rose A Sevcik. 2005. Augmentative communication and early intervention: Myths and realities. *Infants & Young Children*, 18(3):174–185.

Leen Sevens. 2018. *Words divide, pictographs unite: Pictograph communication technologies for people with an intellectual disability*. Netherlands Graduate School of Linguistics.

Leen Sevens, Vincent Vandeghinste, Ineke Schuurman, and Frank Van Eynde. 2015. Extending a Dutch text-to-pictograph converter to English and Spanish. In *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, pages 110–117, Dresden, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Vincent Vandeghinste, Ineke Schuurman Leen Sevens, and Frank Van Eynde. 2017. Translating text into pictographs. *Natural Language Engineering*, 23(2):217–244.

Céline Vaschalde, Pauline Trial, Emmanuelle Esperança-Rodier, Benjamin Lecouteux, and Didier Schwab. 2018. Automatic pictogram generation from speech to help the implementation of a mediated communication. Research report, LIG ; UGA (Université Grenoble Alpes).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2019. Sense vocabulary compression through the semantic knowledge of WordNet for neural word sense disambiguation. In *Proceedings of the 10th Global Wordnet Conference*, pages 108–117, Wroclaw, Poland. Global Wordnet Association.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

8. Language Resource References

André, Virginie and Canut, Emmanuelle. 2010. *TCOF : Traitement de Corpus Oraux en Français*. distributed via Ortolang: https://hdl.handle.net/11403/tcof.

Baude, Olivier and Dugua, Céline. 2008. *ESLO*. also distributed via ORTOLANG: https://hdl.handle.net/11403/eslo. PID http://eslo.huma-num.fr/.

Christophe Benzitoun and Jeanne-Marie Debaisieux. 2020. Orféo: un corpus et une plateforme pour l'étude du français contemporain. *Langages*, (219):160–p.

Benzitoun, Christophe and others. 2016. *CEFC*. distributed via ORTOLANG: https://hdl.handle.net/11403/cefc-orfeo.

Giraudel, Aude and others. 2012. *REPERE Evaluation Package*. distributed via ELRA: ELRA-E0044, ISLRN 360-758-359-485-0.

Gravier, G. and others. 2004. *ESTER Evaluation Package*. distributed via ELRA: ELRA-E0021, ISLRN 110-079-844-983-7.

Gravier, Guillaume and others. 2012. *ETAPE Evaluation Package*. distributed via ELRA: ELRA-E0046, ISLRN 425-777-374-455-4.

Vial, Loïc. 2019. *French Word Sense Disambiguation with Princeton WordNet Identifiers*. distributed via Zenodo: https://doi.org/10.5281/zenodo.3549806.