A Concept Based Approach for Translation of Medical Dialogues into Pictographs

Johanna Gerlach[1], Pierrette Bouillon[1], Jonathan Mutal[1] and Hervé Spechbach[2]

[1] TIM/FTI, University of Geneva, Geneva, Switzerland

[2] HUG, Geneva University Hospitals, Geneva, Switzerland

{johanna.gerlach, pierrette.bouillon, jonathan.mutal}@unige.ch

herve.spechbach@hcuge.ch

Abstract

Pictographs have been found to improve patient comprehension of medical information or instructions. However, tools to produce pictograph representations from natural language are still scarce. In this contribution we describe a system that automatically translates French speech into pictographs to enable diagnostic interviews in emergency settings, thereby providing a tool to overcome the language barrier or provide support in Augmentative and Alternative Communication (AAC) contexts. Our approach is based on a semantic gloss that serves as pivot between spontaneous language and pictographs, with medical concepts represented using the UMLS ontology. In this study we evaluate different available pre-trained models fine-tuned on artificial data to translate French into this semantic gloss. On unseen data collected in real settings, consisting of questions and instructions by physicians, the best model achieves an F0.5 score of 86.7. A complementary human evaluation of the semantic glosses differing from the reference shows that 71% of these would be usable to transmit the intended meaning. Finally, a human evaluation of the pictograph sequences derived from the gloss reveals very few additions, omissions or order issues (<3%), suggesting that the gloss as designed is well suited as a pivot for translation into pictographs.

Keywords: pictographs, medical communication, pre-trained models, UMLS

1. Introduction

Understanding medical information or instructions can be difficult, especially for patients with limited health literacy. Pictographs can facilitate communication in the medical context and have notably been used successfully to deliver specific medical instructions. The use of images has been shown to positively affect patient comprehension by improving attention, recall, satisfaction, and adherence (Houts et al., 2006; Katz et al., 2006; Hill et al., 2016).

Although the potential of pictographs has often been recognised, tools that automatically translate sentences into pictographs are still scarce due to the lack of resources. Some online medical translators include pictographs, for example, “My Symptoms Translator” (Alvarez, 2014) and “Medipicto AP-HP”, but they remain very limited in coverage and can only translate predefined sentences. Pictograph output can also be produced by several generic MT systems, such as Text2Picto (Sevens, 2018; Vandeghinste et al., 2015) and, more recently, PictoBERT (Pereira et al., 2022). Both of these are based on WordNet (Miller, 1995) and provide word-based mapping into pictographs. Glyph (Bui et al., 2012) has improved on the word-based approach in order to build a system for the medical domain. It uses the UMLS (Unified Medical Language System (Bodenreider, 2004) ontology to identify medical terms, in addition to natural language processing and computer graphics techniques. Pictographs are then linked to UMLS concepts rather than words.

Following the same UMLS concept-based approach, we aim to build a system that can translate French speech into pictographs to help French-speaking doctors perform diagnostic interviews in emergency settings, when they do not have a common language with the patient and interpreters are not available, or in Augmentative and Alternative Communication (AAC) contexts where pictographs can improve understanding. To this end, we have built on the resources developed for BabelDr, a speech-enabled fixed-phrase medical translator (Bouillon et al., 2021). While BabelDr relied on neural methods to map spoken interactions to the closest pre-translated target language sentence, our new system, PictoDr, aims to translate these source variations into pictograph sequences using a two-step process: first, spoken or written doctor utterances (questions and instructions) are translated from French text into a UMLS-based semantic gloss, second, this gloss is transposed into pictographs.

The aim of this paper is twofold. First, we want to compare different neural approaches for translation into the semantic gloss and secondly, to evaluate if the design of the semantic gloss is well suited to be a pivot for translation into pictographs. The remainder of this paper is organised as follows: Section 2 describes the PictoDr system in more

LREC-COLING 2024, pages 233–242

20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

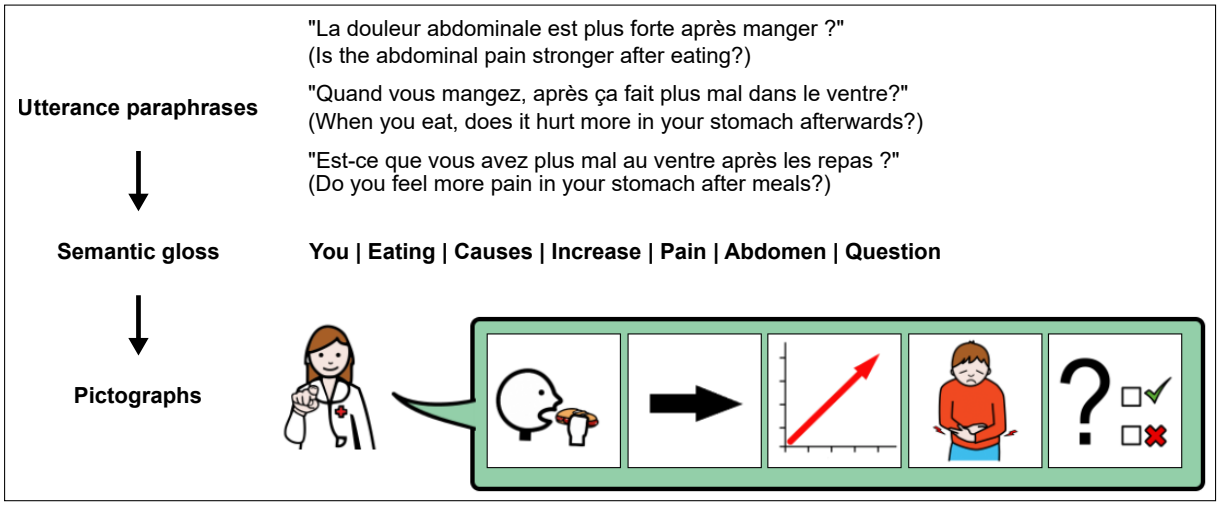

Figure 1: Example of the two-step translation process from French utterance paraphrases to pictographs

detail; Section 3 presents the different MT systems developed for automatic translation of French utterances into the semantic gloss; Section 4 discusses the methodology and results of the evaluation used to compare the approaches for translation into the semantic gloss; Section 5 presents the evaluation of the translations into pictographs and section 6 concludes.

2. PictoDr

To translate French speech into pictographs, PictoDr uses a two-step approach with a semantic gloss serving as pivot, as illustrated in Figure 1.[1] Glossing has been used in many NLP applications, for example, in sign language machine translation (MT), where it is employed to define manual signs and their syntax (Ebling, 2016). It has also been used as a pivot for NMT of low-resourced languages (Zhou et al., 2019). For our purposes, this approach has many advantages. When dealing with paraphrases of the same medical question or instruction, we can map them all to the same gloss and consequently represent them with the same sequence of pictographs. The use of UMLS instead of WordNet also allows us to work with medical terms instead of individual word senses, thereby facilitating the illustration of compounds such as "blood draw", which cannot be adequately rendered by separate illustrations of each word (Norré et al., 2022). Finally, this pivot representation also makes it possible to easily generate different pictograph languages that can be adapted to the language or culture of the patient.

[1]The pictographic symbols used are the property of the Government of Aragón and have been created by Sergio Palao for ARASAAC (https://arasaac.org/), that distributes them under Creative Commons License BY-NC-SA.

In the first step of our translation approach, the result of speech recognition is mapped to the gloss using neural machine translation methods. Training data for this specific task are generated from a manually defined Synchronous Context Free Grammar (SCFG, Aho and Ullman, 1969) that maps source variations, described in a formalism similar to regular expressions, to UMLS semantic gloss, using variables for a compact definition of similar sentences. The semantic gloss consists of an ordered sequence of concepts, combining UMLS concepts such as findings, diagnostic procedures, etc. with non-UMLS functional concepts that indicate agents ("You" in the example in Figure 1) or utterance modes ("Question"). These concepts are ordered based on semantic patterns (for example, <you/patient> <Sign or symptom> <time> <question>). This grammar was defined in close collaboration with physicians from the emergency department of the Geneva University Hospitals (HUG) who helped outline the necessary system coverage, which currently includes all the essential questions and instructions used in emergency triage. The physicians also contributed to the collection of possible spoken variations, i.e. the different paraphrases used for each utterance, as well as their grouping and association with the corresponding gloss. Figure 2 provides an example of a grammar rule.

The current version of the grammar includes 3,096 rules, which expand into 15,488 UMLS glosses once variables are replaced by values. These UMLS glosses are mapped to hundreds of millions of surface variations. To create data for training or fine-tuning, we use this grammar to generate parallel data consisting of source variations (paraphrases) aligned with the corresponding semantic gloss. The generation process consists in expanding all the source variation patterns, replacing

```
{
      "paraphrases": [
        "vous êtes-vous (blessé|blessée) (au niveau du| sur (le|votre)|au) visage",
        (did you injure ?(yourself) (in the area of|on) (the|your) face)
        "?(est-ce que) vous vous êtes (blessé|blessée) (au niveau du| sur (le|votre)|au) visage",
        (did you injure ?(yourself) (in the area of|on) (the|your) face)
        "$avez_vous (une blessure|des blessures) ?(quelque part) (au niveau du| sur (le|votre)|au) visage"
        (did you suffer (an injury|injuries) ?(somewhere) in (the|your) facial area)
      ],
      "umlsGloss": [
        {"semType":"Agent", "concept":"You", "cui":""},
        {"semType":"Finding", "concept":"Past history of", "cui":"C0332119"},
        {"semType":"Injury or Poisoning", "concept":"Injury wounds","cui":"C0043250"},
        {"semType":"Body Location or Region", "concept":"Face", "cui":"C0015450"},
        {"semType":"Mode", "concept":"Question", "cui":""}
      ]
}
```

Figure 2: Example of a grammar rule describing an utterance with its source paraphrases and semantic gloss

variables and then filtering the results using a n-gram based approach as described in (Mutal et al., 2020). For this study, the generated corpus was split into training and validation sets, described in Table 1.

Data	#Sentences	#Words	#Vocabulary
Train.	600K	5M-3M	5,124-1,683
Valid.	17K	138k-92k	2,381-1,683

Table 1: Number of sentences, words and vocabulary in training and validation data sets.

In the second step of our translation approach, the semantic gloss is mapped into a pictograph sequence. In previous work we have created a database linking the concepts used in the gloss to ARASAAC[2] pictographs, where available. This provided illustrations for 66% of the concepts (Gerlach et al., 2023). For the remainder, we have designed new pictographs or adapted existing pictographs. Previous experiments have shown that a plain sequential display of pictographs is not easy to understand, in particular in medical dialogues where multiple actors are involved. A contextualised display approach is therefore used. Different visualisations are provided depending on the patterns, for example, for yes-no questions, wh-questions, instructions, etc. Additionally, in this step, adjacent UMLS concepts such as “Lesion” and “Skin”, which are difficult to represent individually, can be replaced by complex ones (“Skin lesions”).

3. MT systems

The aim of the machine translation system is to translate French speech transcriptions into a semantic gloss. The same architecture (standard Transformer encoder-decoder (Vaswani et al., 2017)) and same validation and training data were used for all the approaches. Our baseline system uses the same settings as described in (Mutal et al., 2022):

Baseline. The Baseline was trained from scratch with our artificial data using OpenNMT (Klein et al., 2017), involving 200,000 steps with early stopping criteria. Our training process resulted in a 99% perplexity on the validation dataset.
Since the source in this task is a high-resource language, we compare this Baseline with different systems leveraging available pre-trained encoders. However, for the target, which is an artificial semantic gloss, no pre-trained models are available. Therefore for all models we used the same decoder (L=16, H=512, A=16) with a vocabulary size equivalent to that of the target side of the training data (semantic gloss tokens), augmented by special tokens. We fine-tune these models for our task using the artificial data described above (cf. Table 1).
We experimented with four different BERT-based (Devlin et al., 2019) pre-trained encoders that we will now describe in further detail:

CamemBERT. CamemBERT[3] (Martin et al., 2020) is based on RoBERTA (Zhuang et al., 2021). It uses the French part of the OSCAR corpus (Open Super-large Crawled Aggregated corpus) (Ortiz Suárez et al., 2019) for pre-training.

FlauBERT. FlauBERT[4] is pre-trained on a French corpus consisting of sub-corpora covering diverse topics and writing styles, from well-written

[2] https://arasaac.org/

[3] https://huggingface.co/camembert-base

[4] https://huggingface.co/flaubert/flaubert_base_uncased

text (e.g. Wikipedia and books) to Common Crawl (Le et al., 2020).

DrBERT. DrBERT,[5] a bio-medical model, follows the same architecture and pre-training strategy as CamemBERT. The model was only pre-trained on a small data set from the healthcare domain (Labrak et al., 2023). We chose this model to see if domain-specific pre-training improves performance on the task (Lee et al., 2020).

XML-R. XML-R[6] is a multilingual version of RoBERTa pre-trained on 100 languages (Conneau et al., 2020). We included this model on the assumption that a multilingual approach could make the system more robust to medical terminology of different origins (e.g. anglicisms) in the source.

All these models were trained using the HuggingFace (Wolf et al., 2020) framework. We used default hyper-parameters for each approach. The number of parameters, vocabulary and pre-training data for each system are shown in Table 2.

Model	#Params	#Vocab.	Data
Baseline	196M	32k	-
CamemBERT	196M	32k	138 GB
FlauBERT	224M	50k	71GB
DrBERT	196M	32k	7 GB
XML-R	360M	250k	2.5TB

Table 2: Number of parameters for the encoder-decoder, vocabulary size and size of pre-training data measured in bytes.

4. Evaluation of translation into semantic gloss

The aim of this first evaluation is to see how well the different models can translate French utterances into the semantic gloss. In particular, we would like to see 1) whether pre-trained models are of use for this particular task, and 2) whether the amount and type of pre-training data (domain specific, multilingual, etc) have an impact on system performance.

In a first automatic evaluation, we compare system outputs against references using $F0.5$ and $F0.5_{3-best}$ scores. In a second human evaluation that takes into consideration only the best-performing system according to the first evaluation, we assess whether the outputs that do not match the references can still be considered correct or useful by doctors. In the following sections we describe the data, followed by these two evaluations.

4.1. Evaluation Data

We used data collected with the BabelDr application at the HUG during diagnostic interviews in real triage settings. These speech data were transcribed manually and each utterance was manually annotated with a reference gloss following the same patterns as for the training data (cf. Section 2). The data are divided into two sets, a first larger set (HUG1) that was used to improve the coverage of the grammar and define the semantic gloss, and a second smaller set (HUG2) consisting of unseen data.[7] An overview of these two sets is provided in Table 3.

	HUG1	HUG2
#Sentences	1,252	380
#OOT	993	320
#IT	259	60
#Words	7,670-4,735	2,441-1,346
#Vocabulary	895-408	440-207
#Vocab. Coverage	93%	91%

Table 3: Number of sentences, out of training sentences (*OOT*), in training sentences (*IT*), number of words/gloss tokens, size of vocabulary and vocabulary coverage (ratio of the vocabulary covered by the data used to specialise the models) for the test sets.

4.2. Automatic evaluation

As both precision and recall are important for the task, we chose to use the F-measure for this evaluation. Since our focus is on precision, we use a weighted F-measure, with $\beta = 0.5$ ($F0.5$) to give a higher weight to precision. The definitions of true positive, false positive and false negative used for this metric are provided in Table 4.

The models generate n-best results, which can be of use for the task and are therefore displayed in the application for the user to choose from. In order to assess their usefulness, we calculated a second F0.5 score ($F0.5_{3-best}$) which takes into account the top three outputs by calculating a weighted average. Based on the assumption that users would not need to look beyond the first suggested result if it served their purposes, a higher weight was assigned to the first output, as illustrated in the following equation:

[5] https://huggingface.co/Dr-BERT/DrBERT-7GB

[6] https://huggingface.co/xlm-roberta-base

[7] The data from this second corpus including the different forms used in this study (transcription, reference semantic gloss and pictograph sequence) is available here https://propicto.demos.unige.ch/lrec2024/

Type	Definition	Example
Transcription	Physician's utterance	Combien de fois avez-vous vomi ? (How many times did you vomit?)
Hypothesis	System output	[Times/day] [You] [Vomiting] [Question]
Reference	Reference semantic gloss	[How many times did this happen] [You] [Vomiting] [Question]
True Positives (TP)	Number of correct gloss tokens in the hypothesis	[You] [Vomiting] [Question]
False Positives (FP)	Number of additional gloss tokens in the hypothesis	[Times/day]
False Negatives (FN)	Number of missing gloss tokens in the hypothesis	[How many times did this happen]
Formula	$F0.5 = \frac{1.25*TP}{1.25*TP+0.25*FN+FP}$	$F0.5 = \frac{1.25*3}{1.25*3+0.25*1+1} = 0.75$

Table 4: Definition of True Positive, False Negative and False Positive. The differences between the hypothesis and reference are marked in red.

$$F0.5_{3-best} = 0.7 * F0.5(Output_{1-best}, Ref) + 0.2 * F0.5(Output_{2-best}, Ref) + 0.1 * F0.5(Output_{3-best}, Ref) \quad (1)$$

Results for all evaluated models are shown in Table 5. Overall, the machine translation systems using pre-trained models consistently outperformed the baseline, achieving an improvement of at least 20 points in both $F0.5$ and $F0.5_{3-best}$. Among the different pre-trained models the difference in $F0.5$ scores was small, indicating that these systems generated similar first output with high $F0.5$ scores, with the best model achieving 94.73% on HUG1 and 86.71% $F0.5$ on HUG2.

Overall, CamemBERT outperformed the other pre-trained models by a small margin. XML-R demonstrated the second-best performance when only taking into consideration the first output ($F0.5$). However, when taking into account the 3-best ($F0.5_{3-best}$), DrBERT achieved better results than its counterparts, suggesting that the n-best results have more elements in common with the reference and might be more useful for the task.

As is often observed in the literature, the domain of the data used for pre-training has an impact on performance on the task (Lee et al., 2020). For our task, the medical pre-training data used for DrBERT, despite being a much smaller quantity, allows this model to be competitive with other models pre-trained on much larger data sets (e.g. CamemBERT) or using larger numbers of parameters (e.g. XML-R).

4.3. Human evaluation

In this complementary evaluation of the unseen data test set (HUG2), we asked two physicians from the HUG emergency department to evaluate the sentences where the system output was different from the reference (N=56). Participants were shown the original transcription along with the output (semantic gloss) of the best system according to the automatic evaluation (CamemBERT) and were asked to select one of the following options: "same meaning", "similar meaning" and "different meaning". Instructions for the task defined similar meaning as "has a close meaning that could also be of use in the diagnostic interview to obtain the required information from the patient". In case of doubt, participants could select a "I don't know" option.

A breakdown of the results from this human evaluation is presented in Table 6. The evaluators reached the same judgement for a large share of the sentences (80%). According to Cohen's kappa ($\kappa = 0.661$), the level of agreement is moderate. Taking into account only the items where the two physicians agree on either "same meaning" or "similar meaning", results suggest that 71% of the system outputs that were discarded in the automatic evaluation because they do not match the reference exactly can still be of use in a diagnostic interview. Table 7 provides examples of sentences of each category. The "same meaning" items are mostly paraphrases or very close meanings that serve the same purpose (e.g. "sport" vs "exertion"). The "similar meaning" cases are often sentences where the gloss is either more precise or more generic than the transcription ("abdominal pain" vs "pain"). The "different meaning" cases are

	HUG1		HUG2	
Model	$F0.5$	$F0.5_{3-best}$	$F0.5$	$F0.5_{3-best}$
Baseline	69.71	67.14	67.63	64.26
CamemBERT	94.32	86.79	86.71	79.15
FlauBERT	91.73	84.51	84.39	77.75
DrBERT	93.12	86.11	85.11	78.49
XML-R	93.79	82.73	85.94	75.17

Table 5: $F0.5$ and $F0.5_{3-best}$ between the system outputs and the reference. The scores were calculated on our two test sets.

mostly associated with out of coverage structures or terms (e.g. drug names not included in the training data, such as “Tramadol”).

Meaning	Eval.1	Eval.2	Agree
Same	32	32	29
Similar	13	19	11
Different	10	5	5
I don’t know	1	-	-
Total	56	56	45

Table 6: Human evaluation of the gloss: results by category for the two physicians

5. Evaluation of translation into pictographs

The aim of this second evaluation is to see whether the proposed approach (using a semantic gloss as pivot) produces usable pictograph sequences that are perceived as conveying the same meaning as the original utterance, despite the abstraction and simplification introduced by the pivot. To this end, we present French speaking participants with utterances paired with their final pictograph representations.

5.1. Test data

For this evaluation, only the second smaller data set (HUG2) was used. As this corpus is composed of multiple diagnostic interviews, some common utterances (e.g. “what brings you to the emergency department?”) occur many times, with different paraphrases in French, but resulting in the same semantic gloss and thus the same pictograph representation. To reduce the number of items to evaluate, for each unique gloss included in the corpus, only the first source paraphrase was retained. We also chose to remove incomplete sentences that, due to the lack of dialogue context, were ambiguous in terms of meaning and thus not suited for our evaluation. The resulting data set included 147 French utterances with corresponding semantic glosses. Examples of utterances with their glosses and representations are provided in Figure 3.

5.2. Design

To avoid bias introduced by model errors, the pictograph sequences were generated from the reference glosses rather than system output. Participants were shown source sentence and pictograph translations utterance by utterance in a LimeSurvey, and were asked to rate these by means of the following multiple choice options: “the translation is correct”, “one or more pictographs are unnecessary”, “one or more pictographs are missing”, “one or more pictographs are incorrect”, “I don’t understand one or more pictographs”, “the order of the pictographs is not logical”. For all choices, participants were asked to indicate in a free text field which of the individual pictographs were concerned.

The data set containing 147 utterance-gloss pairs was divided into four smaller subsets consisting of 37, 37, 37 and 36 items respectively. We recruited a total of 16 participants, allowing us to collect four responses for each subset. Participants were recruited at the Faculty of translation and interpreting and are either native or near-native speakers of French.

5.3. Results

The results of this evaluation are reported in Table 8. Overall, more than half (57%) of the pictograph sequences were found to be appropriate translations of the source sentences. No majority was reached for nearly a quarter of the evaluated items, suggesting that the task is highly subjective. Inter-annotator agreement for this question is low in terms of observed agreement (47.6%) and fair according to Fleiss’ Kappa ($\kappa = 0.324$). A closer analysis of the 31 sentences that a majority of participants judged as incorrect shows that the most important issue (N=12) were pictographs that

Judgement	Transcription	Semantic gloss
same meaning	Avez-vous mal seulement quand vous faites un effort physique ? (Do you only feel pain when you exert yourself?)	You | Sports | Causes | Pain | Question
similar meaning	Je vais vous prescrire des médicaments laxatifs (I will prescribe you laxatives)	I | Prescription procedure | Pharmaceutical Preparations | Gastrointestinal transit function
similar meaning	Depuis quand vous avez des douleurs au ventre ? (Since when do you have abdominal pain?)	Since when | You | Pain | Question
similar meaning	Est-ce que la douleur est comme un poids ? (Is the pain like a heaviness?)	You | Stomach feeling heavy | Question
different meaning	Est-ce que le tramadol a diminué la douleur ? (Did the Tramadol relieve the pain?)	You | Toradol | Causes | Positive | Question

Table 7: Human evaluation of the gloss: examples

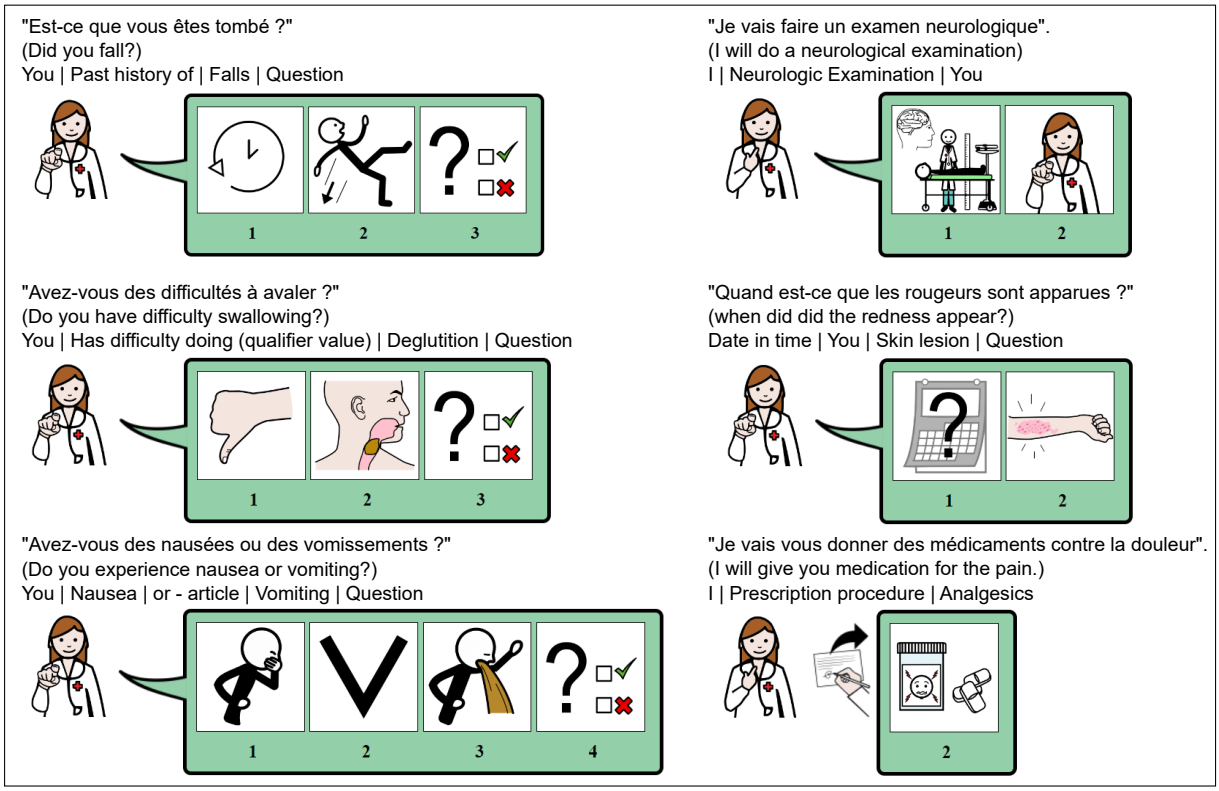

Figure 3: Examples of utterances with corresponding semantic gloss and pictograph sequence included in the second evaluation

were not understandable, often for medical concepts (e.g. “Neurologic Examination”) or for time indications (e.g. “Recent”).

For the questions related to the composition of the pictograph sequence (additions, omissions, incorrect pictographs and bad order), the observed agreement is much higher (82-85%). Results show a very low proportion of perceived omissions, incorrect pictographs or order issues (1% each). Additions are slightly more frequent (3%). The representation of time is responsible for a large share of the reported additions, as in some utterances the time frame is implicit and does not require illustration (e.g. “did you fall?”, which im-

plicitly happened in the past), whereas in others the temporal condition requires a precise illustration (e.g. “have you recently had an operation?”). Other pictograms that were found unnecessary were those indicating a causal relationship in the context of another causal concept (e.g. in “Eating | Causes | Increase | Pain”) or disjunctions (e.g. “Nausea | or - article | Vomiting”). As for the order of the pictographs, it was reported as illogical in very few cases, mainly for pairs of pictograms including a qualifier and an action, such as “Has difficulty doing” + “Deglutition”.
These results show that the concepts included in the gloss are mostly appropriate, although they are not always well illustrated by the current set of pictographs.

	Yes	No	No maj.
Correct	84 (57%)	31 (21%)	32 (22%)
Addition	4 (3%)	133 (90%)	10 (7%)
Omission	1 (1%)	142 (97%)	4 (3%)
Bad picto.	1 (1%)	144 (98%)	2 (1%)
Incompr.	9 (6%)	128 (87%)	10 (7%)
Bad order	2 (1%)	139 (95%)	6 (4%)

Table 8: Results of the evaluation of the pictograph sequences showing majority judgements for the 147 sentences

6. Conclusions and future work

In a low-resourced domain, this study evaluates the feasibility of using a UMLS-based semantic gloss as pivot in a two-step approach to translate speech into pictographs.
For the first step, translating from a high-resourced language, the evaluated pre-trained models outperformed a baseline trained from scratch. Overall the different models give comparable results, with CamemBERT slightly outperforming the others. While these approaches work well in the limited domain studied here, future work should investigate methods to make the systems more robust by including other medical domains. However, while the grammar (and the artificial training data generated from it) could easily be extended, the expansion of system coverage remains limited by the scarce availability of medical pictographs. Pictographs designed for AAC are typically limited to simple and concrete concepts, since the target audience must possess prior knowledge of these concepts and how they are represented for the pictographs to be effective. Therefore, illustrations of body parts or tools used for exams are readily available, while more complex or abstract concepts are absent from existing sets. Another issue is the the lack of metadata associating pictographs with medical concepts.
Regarding the second translation step, a human evaluation of pictograph sequences derived from the gloss has shown that the abstraction from the original surface form introduced by the pivot is not perceived as a distortion of meaning. The use of concept sequences that sometimes bear little resemblance to the original words and their order, as well as the simplification accomplished by keeping only essential semantic elements were only very rarely reported as problematic by evaluators. The main issue we observed is the complexity of illustrating some of the concepts that are important in the medical domain. These results suggest that the semantic gloss as designed is well suited as a pivot for translation into pictographs.
The current system will serve as basis for further more detailed evaluations of sentence patterns and pictograph order, in particular for patients from different cultures and languages. In terms of future work, we also plan on leveraging the concepts and pictographs from the current system to build an interface which would allow patients to describe their symptoms or medical history by selecting pictographs and combining them in patterns, which could then be translated into French.
The system is available online at the address https://propicto.demos.unige.ch/pictoDrClient.

7. Acknowledgements

This work is part of the PROPICTO project, funded by the Swiss National Science Foundation (N°197864) and the French National Research Agency (ANR-20-CE93-0005).

8. References

Alfred V. Aho and Jeffrey D. Ullman. 1969. Syntax directed translations and the pushdown assembler. *Journal of Computer and System Sciences*, 3(1):37–56.

Juliana Alvarez. 2014. Visual design. A step towards multicultural health care. *Arch Argent Pediatr*, 112(1):33–40.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(90001):267D – 270.

Pierrette Bouillon, Johanna Gerlach, Jonathan Mutal, Nikos Tsourakis, and Hervé Spechbach. 2021. A speech-enabled fixed-phrase translator for healthcare accessibility. In *Proceedings of the 1st Workshop on NLP for Positive Impact*,

pages 135–142, Online. Association for Computational Linguistics.

Duy Duc An Bui, Carlos Nakamura, Bruce E. Bray, and Qing Zeng-Treitler. 2012. Automated illustration of patients instructions. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1158. American Medical Informatics Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sara Ebling. 2016. *Automatic Translation from German to Synthesized Swiss German Sign Language*. Thesis for the degree of Doctor in Philosophy, University of Zurich.

Johanna Gerlach, Pierrette Bouillon, Magali Norré, and Hervé Spechbach. 2023. Translating medical dialogues into pictographs: An approach using UMLS. In *Caring is Sharing - Exploiting the Value in Data for Health and Innovation - Proceedings of MIE 2023, Gothenburg, Sweden, 22 - 25 May 2023*, volume 302 of *Studies in Health Technology and Informatics*, pages 823–824. IOS Press.

Brent Hill, Seneca Perri-Moore, Jinqiu Kuang, Bruce E Bray, Long Ngo, Alexa Doig, and Qing Zeng-Treitler. 2016. Automated pictographic illustration of discharge instructions with Glyph: impact on patient recall and satisfaction. *Journal of the American Medical Informatics Association*, 23(6):1136–1142.

Peter S. Houts, Cecilia C. Doak, Leonard G. Doak, and Matthew J. Loscalzo. 2006. The role of pictures in improving health communication: A review of research on attention, comprehension, recall, and adherence. *Patient Education and Counseling*, 61(2):173–190.

Marra G. Katz, Sunil Kripalani, and Barry D. Weiss. 2006. Use of pictorial aids in medication instructions: A review of the literature. *American journal of health-system pharmacy*, 63(23):2391–2397.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, page 67–72, Vancouver, Canada. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. Drbert: A robust pre-trained model in french for biomedical and clinical domains. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL'23), Long Paper*, Toronto, Canada. Association for Computational Linguistics.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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.

Jonathan Mutal, Pierrette Bouillon, Johanna Gerlach, Paula Estrella, and Hervé Spechbach. 2019. Monolingual backtranslation in a medical speech translation system for diagnostic interviews - a NMT approach. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 196–203,

Dublin, Ireland. European Association for Machine Translation.

Jonathan Mutal, Pierrette Bouillon, Magali Norré, Johanna Gerlach, and Lucia Ormaechea Grijalba. 2022. A neural machine translation approach to translate text to pictographs in a medical speech translation system - the BabelDr use case. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 252–263, Orlando, USA. Association for Machine Translation in the Americas.

Jonathan Mutal, Johanna Gerlach, Pierrette Bouillon, and Hervé Spechbach. 2020. Ellipsis translation for a medical speech to speech translation system. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 281–290, Lisboa, Portugal. European Association for Machine Translation.

Magali Norré, Vincent Vandeghinste, Pierrette Bouillon, and Thomas François. 2022. Investigating the Medical Coverage of a Translation System into Pictographs for Patients with an Intellectual Disability. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 44–49. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut f"ur Deutsche Sprache.

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

Leen Sevens. 2018. *Words Divide, Pictographs Unite: Pictograph Communication Technologies for People with an Intellectual Disability*. LOT, JK Utrecht, The Netherlands.

Vincent Vandeghinste, Ineke Schuurman, Leen Sevens, and Frank Van Eynde. 2015. Translating text into pictographs. *Natural Language Engineering*, 23(2):217–244.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, page 38–45, Online. Association for Computational Linguistics.

Zhong Zhou, Lori S. Levin, David R. Mortensen, and Alexander H. Waibel. 2019. Using interlinear glosses as pivot in low-resource multilingual machine translation. *arXiv: Computation and Language*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.