A Speech-enabled Fixed-phrase Translator for Healthcare Accessibility

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

[1]FTI/TIM, University of Geneva, Switzerland
[2]Hôpitaux Universitaires de Genève (HUG), Switzerland
{Pierrette.Bouillon, Johanna.Gerlach, Jonathan.Mutal, Nikolaos.Tsourakis}@unige.ch
Herve.Spechbach@hcuge.ch

Abstract

In this overview article we describe an application designed to enable communication between health practitioners and patients who do not share a common language, in situations where professional interpreters are not available. Built on the principle of a fixed phrase translator, the application implements different natural language processing (NLP) technologies, such as speech recognition, neural machine translation and text-to-speech to improve usability. Its design allows easy portability to new domains and integration of different types of output for multiple target audiences. Even though BabelDr is far from solving the problem of miscommunication between patients and doctors, it is a clear example of NLP in a real world application designed to help minority groups to communicate in a medical context. It also gives some insights into the relevant criteria for the development of such an application.

1 Motivation

Access to healthcare is an important component of quality of life, but it is often compromised by the language barrier which prevents effective communication. In hospitals, medical staff are increasingly confronted with patients with whom they do not share a common language. Lack of clear communication can lead to increased risk for patients (Flores et al., 2003) but also discourages vulnerable groups from seeking medical assistance. When professional interpreters are not easily available, for example in emergency situations, there is a crucial need for tools to overcome the language barrier in order to provide medical care. While many generic translation solutions are available on the web, they present numerous disadvantages, including the unreliability of machine translation (Bouillon et al., 2017), the insufficient data confidentiality of cloud services or the absence of resources for minority languages. To overcome these issues, specifically designed tools based on a limited set of pre-translated sentences have been developed. These phraselators (Seligman and Dillinger, 2013) have the advantage of portability, accuracy and reliability. Although these tools have limited coverage, and do not solve all communication issues, recent studies have shown that they are generally preferred to machine translation as they are perceived as more reliable and trustworthy in these safety critical contexts (Panayiotou et al., 2019; Turner et al., 2019).

This paper aims to provide an overview of the NLP components included in the speech-enabled phraselator called BabelDr. In Section 2 we will give an overview of BabelDr usage. We then explain the artificial training data derived from the grammar to specialise the different components in Section 3. In sections 4, 5, 6, 7 and 8 we explain BabelDr's components in detail, as well as the possible outputs available to users. We then present several usage studies with target groups in Section 9.1, report on the performance of the whole system in Section 9.2 and conclude in Section 10.

2 BabelDr

BabelDr[1] is a joint project between the Faculty of Translation and Interpreting of the University of Geneva and Geneva University Hospitals (HUG). (Bouillon et al., 2017). The aim of the project is to develop a speech to speech translation system for emergency settings which meets three criteria: reliability, data security and portability to low-resource target languages relevant for HUG. It is designed to allow French-speaking medical practitioners to carry out triage and diagnostic interviews with patients speaking Albanian, Arabic, Dari, Farsi, Spanish, Swiss French sign language and Tigrinya.

[1]More information available at https://babeldr.unige.ch/

Proceedings of the 1st Workshop on NLP for Positive Impact, pages 135–142
Bangkok, Thailand (online), August 5, 2021.

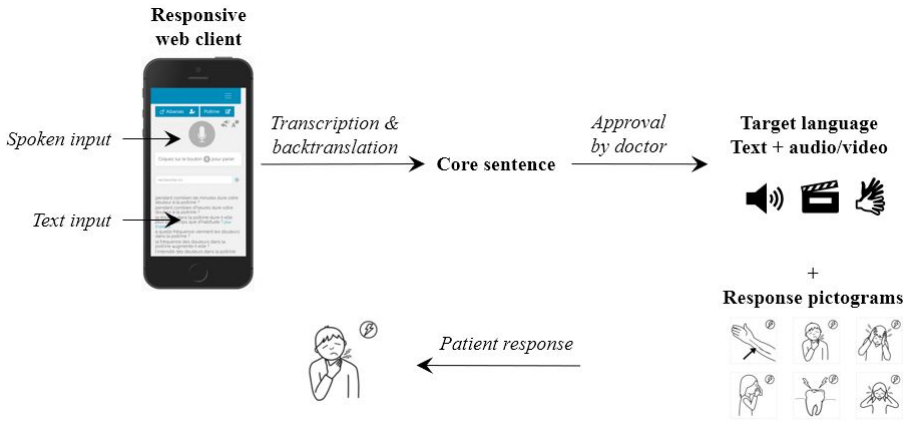

Figure 1: Overview of BabelDr usage

BabelDr is a web application designed to function on desktops and mobiles. Built on the principle of a phraselator, it relies on a limited set of pre-translated sentences, hereafter called core-sentences, collected with doctors. For improved usability and more natural interaction with the patient, it includes a speech recognition component: instead of searching for utterances in menus, medical staff can speak freely and the system will map the spoken utterances to the closest pre-translated core-sentence. This sentence is then presented for validation, in a backtranslation step, ensuring that the doctor knows exactly what is being translated for the patient. The patient can then respond by means of a pictogram-based interface. All components can be deployed on a local server with no dependency on cloud services, thus ensuring the data confidentiality that is essential for medical applications. Figure 1 illustrates the usage of BabelDr.

3 Training data and grammars

Due to confidentiality issues, training data for spoken French medical dialogues is scarce. For this reason, the first version of the system was built around a manually defined Synchronous Context Free Grammar (SCFG, Aho and Ullman, 1969), used for grammar-based speech recognition and parsing (Rayner et al., 2017). This grammar is now leveraged to generate artificial data used both for backtranslation (Section 5) and for specialising speech recognition (Section 4).

The grammar maps source variation patterns, described in a formalism similar to regular expressions, to core-sentences. Due to the repetitive nature of the content, the grammars make use of compositional sentences to make resources more compact. These sentences contain one or more variables, which are replaced by different values at system compile time. Figure 2 gives an example of a compositional utterance rule.

The current version of the grammar includes 2629 utterance rules, organised by medical domain, which expand to 10'991 core-sentences once variables are replaced by values. These core-sentences are mapped to hundreds of millions of surface sentences. Figure 3 shows an example of the aligned core-sentence - variation corpus that can be generated from the grammar.

4 Speech-to-Text

To ensure both accuracy and usability, the system uses a hybrid approach for speech recognition, combining two recognisers. The first is a grammar based speech recogniser using GRXMLs generated from the original SCFG (see Section 3). While this is fast and accurate, since it directly yields a core-sentence, it is unable to handle utterances that are out of grammar coverage. It is therefore complemented by a large-vocabulary recogniser specialised with the monolingual artificial corpus described in Section 3. The results of the two approaches are combined based on the confidence score provided by the grammar based recogniser: if the score is over a pre-defined threshold, this result is kept, else the system falls back on the large-vocabulary result. Their performance is evaluated in terms of WER, which is 38.9% for the GRXML grammar and 14.4% (see Table 2) for the large vocabulary model. In this case we have used the dataset of the user study described in (Bouillon et al., 2017).

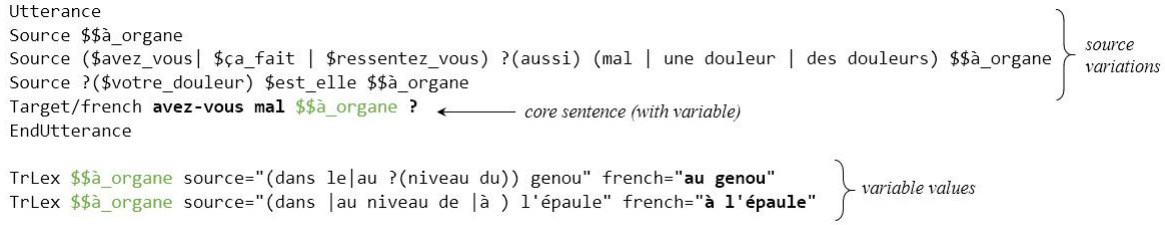

Figure 2: Example of a grammar rule

avez-vous de la fièvre ?;les douleurs étaient-elles accompagnées par de la fièvre
avez-vous de la fièvre ?;vous sentez-vous fiévreux maintenant
avez-vous de la fièvre ?;vous avez une sensation d'être fiévreuse
avez-vous de la fièvre ?;êtes-vous fiévreuse en ce moment
avez-vous de la fièvre ?;sont-elles accompagnées de température
avez-vous de la fièvre ?; est-ce que vous sauriez si vous avez de la fièvre maintenant
avez-vous de la fièvre ?;fébrile
avez-vous de la fièvre ?;avez-vous de la température
avez-vous de la fièvre ?; chaud
avez-vous de la fièvre ?; …….

Figure 3: Example of the aligned corpus generated from the grammar: core-sentences with corresponding source variations

For the GRXML recogniser we use the Nuance ASR v10 and the Nuance Transcription Engine 4 for the large-vocabulary one. Both can be accessed over the network through our custom API using HTTP POST requests. The recognition is file-based and it proves to work well for any real-time interaction. The distributed nature of our back-end platform permits easy scaling and load balancing so that multiple users can interact simultaneously with the recognisers. Especially, for the GRXML case, we can load and compile grammars on the fly or change the parameters of the recogniser dynamically. We can also parse any text against a specific grammar using an HTTP request.

5 Backtranslation

The backtranslation (introduced in Section 2) is an essential step in BabelDr since it maps the speech recognition result to a core-sentence that is presented to the doctor for validation. For the GRXML recogniser, backtranslation is performed directly by the grammar. For the large vocabulary recogniser, as the set of core-sentences is limited (see Section 3), the backtranslation task can be seen as a sentence classification task where the core-sentences are the categories, or as translation task into a controlled language. As a resource, we use the bilingual corpus generated from the grammar as training data. Rayner et al. (2017) introduced an approach based on tf-idf indexing and dynamic programming (DP) achieving 91.8% on accuracy (assuming perfect speech recognition and 1-best). Mutal et al. (2019) then applied different approaches using deep learning methods, neural machine translation (NMT) and sentence classification achieving 93.2% (see Table 2) accuracy on core-sentence matching for transcriptions (assuming perfect speech recognition), improving on the previous approach. This approach is currently used in BabelDr.

6 Elliptical Sentences

In dialogues, elliptical utterances are very common, since they ensure the principle of economy and usually avoid duplication (Hamza, 2019). In BabelDr, they allow doctors to question patients in a more efficient way (Tanguy et al., 2011). However, literal translation of these utterances could affect communication as illustrated in Table 1. In BabelDr, elliptical utterances are not translated literally, but are instead mapped to the closest non-elliptical core-sentence, based on the context.

To avoid a wrong backtranslation in elliptical sentences, a context-level information (the previous accepted utterance) is added to the model. Therefore, when an utterance is identified as an ellipsis,

Utterance	Translation
do you have pain in your stomach?	¿le duele el estómago?
in your head?	*¿en tu cabeza?
Good Translation: ¿Le duele la cabeza?	

Table 1: Example of a bad translation of ellipsis. The * means a bad translation.

it is concatenated with the previous translated utterance before backtranslating. In the context of BabelDr, elliptical utterances are detected using a binary classifier. The model was trained using handcrafted features, such as sentence length, absence of verbs or nouns, part of speech of the first word, and identification of pronouns that refer to entities in the context (using morphological features). On an artificial ellipsis data set, the model achieves 98% accuracy on detecting elliptical sentences and 88% on backtranslating them to a core-sentence (see more, Mutal et al., 2020).

7 Output

After validation of the backtranslation, BabelDr presents the target language output to the patient in written and spoken form, which are both based on the same human translations of the core-sentences. In the following sections we first outline the translation approach and then describe how the translations are rendered for the patient, in audio (for spoken languages) or video format (for sign language).

7.1 Translation

High translation quality is essential for a medical phraselator, therefore the translations are produced by professional translators. Translating for BabelDr presents technical challenges, since language resources must be in a specific structured data format not easily accessible to translators. An online translation platform which includes a translation memory and allows translators to efficiently handle the compositional items was developed to facilitate the translators' task and ensure the quality and coherence of the translations (Gerlach et al., 2018).

The translations are aimed at patients with no medical knowledge and designed to be understandable by patients with a low level of literacy. Sentences were also adapted to account for cultural aspects, such as sensitive or intimate topics that are not commonly discussed, related for example to sexual habits (Halimi et al., 2020). Since the system provides translations both in written and spoken form, the translators had to choose phrasings that would function in both. A recent evaluation of the translations for two of the system's target languages (Albanian and Arabic) has shown that these translations are easy to understand, and thereby make the system more trustworthy in comparison to MT (in publication, Gerlach et al., 2021).

Ongoing developments include the extension of the system to new target languages and modalities to make the system accessible to further population groups. One addition involves translation to pictographs targeted at people with intellectual disabilities, another is translation into easy language, beginning with Simple English.

7.2 Text-to-Speech

Audio has been an important output modality for the BabelDr system, as it presents various competitive advantages for the patients. It alleviates the burden of looking on the screen, which proves to be challenging in a medical setting, e.g. positioning of the physician and patient. Especially, for illiterate users, it is an essential component, and having a system talking in their own language can improve user experience. While it would be possible to have a human record all the pre-translated sentences, due to the number and repetitive nature of the sentences, the time and cost involved in recording were considered too high. The option of a Text-to-Speech (TTS) system was therefore adopted from the beginning of the project in order to announce the translated questions of the physician. State-of-the-art systems like Nuance Vocalizer are now part of our content creation pipeline for crafting the prompts.

Systems of this kind, however, lack support for low-resource languages that the BabelDr system also targets. For this reason, we have investigated the option of building our own TTS for those languages from scratch. In a previous study, positive feedback in terms of comprehensibility was

Figure 4: Doctor and patient interfaces

Task	Model	Metric	Result
Speech to Text	GRXML	WER	38.9%
	Large Vocabulary		14.4%
Back Translation	NMT	Accuracy	93.2%
Overall (3-best)		SER	5%

Table 2: Performance by component and overall

received (Tsourakis et al., 2020), after building a synthetic female voice for the Albanian language based on Tacotron 2, a neural network architecture for speech synthesis directly from text (Shen et al., 2017). Among the target languages supported by BabelDr, Tigrinya is one for which no public TTS is available.

For this reason, a female voice talent was recruited to record all the prompts that were subsequently used in the online system. This allowed us to create a corpus with 18 hours of speech that we exploit in order to create the Tigrinya synthesized voice. The training process is similar to the one found in (Tsourakis et al., 2020). As new content is constantly added to the system, new recordings of the translations are requested. This time we first generate the output with the TTS and ask the voice talent to listen to the prompts. If the result is acceptable the TTS version is kept, otherwise, a human recording is necessary. In a set of 2150 prompts the human had to record 573 files (26.7%).

7.3 French Sign Language

Establishing effective and reliable communication between a doctor and a deaf patient is a complicated task. The scarcity of professional interpreters and the lack of awareness of medical staff for deaf culture severely impedes communication. To create sign language output for our fixed-phrase translator, we have investigated two different approaches: recorded human signers and an avatar (using JASigning, Glauert and Elliott, 2011). An evaluation carried out with the deaf community showed that the recorded human signers are superior in terms of understandability and acceptability, but it was found that the avatar could be useful in this context (in print, Bouillon et al., 2021). The recorded videos were recorded by a sign language interpreter in collaboration with a deaf nurse, and are freely accessible in the online system, providing a human translation reference in sign language for medical questions. These resources present opportunities to evaluate what affects the communication

task with deaf people in this specialised context.

8 Patient response interface

The original BabelDr system was limited to yes-no questions or questions where the patient could respond non-verbally, for example by pointing at a body part. This restrictive approach was problematic both for doctors, who are used to asking open questions, and for patients who had little means to actively contribute to the direction of the dialogue. To build a bidirectional version that would allow more complex responses from the patient, we considered different options. Building a system that would allow patients to respond with speech presents numerous difficulties. No speech recognisers exist for many of the minority languages targeted by our system, and few or no resources such as speech corpora are available to build such systems. A text interface, as found in traditional phraselators, while easier to implement, would not be accessible to patients with low literacy. Additionally, in the context of a fixed phrase translator, some user training is necessary to familiarise with system coverage, which is not possible for patients who arrive at an emergency service. For these reasons, we chose to add a simple pictograph based response interface, shown in Figure 4. Each core-sentence is linked to a set of corresponding response pictographs among which the patient can select their response. Evaluation of these pictographs in terms of understandability and acceptability by patients of different educational and cultural backgrounds is ongoing (Norré et al., 2020). A task-based evaluation showed that all patients prefered the bidirectional version since they could explain their symptoms more efficiently.

9 Evaluation

9.1 Task based

A translation system for the healthcare domain should be evaluated on the task it is designed to assist, which in the case of BabelDr is the diagnostic interview. To this end, we carried out several usage studies. In a preliminary study, we asked four medical students and five doctors to diagnose two standardised Arabic speaking patients, using BabelDr and Google Translate (GT). Results showed that in comparison to the generic machine translation tool, BabelDr provides higher-quality translations and led to a higher number of correct diagnoses (8/9 for BabelDr against 5/9 for GT), in particular with medical students (Bouillon et al., 2017). A subsequent crossover study where 12 French speaking doctors where asked to diagnose two Arabic speaking standardised patients using BabelDr confirmed that the application allows doctors to reach accurate and reliable diagnoses (24/24 correct). It was agreed among participating medical professionals that BabelDr could be used in their everyday medical practice (Spechbach et al., 2019).

The system is currently in use at the HUG outpatient emergency unit and a user satisfaction study is ongoing to collect patients' and doctors' feedback on system usage in real emergency settings by means of questionnaires (Janakiram et al., 2020). The study includes only patients with no understanding of French and no common language with the doctor. Overall, 90% of the 30 patients included so far reported a positive level of satisfaction. The doctors reported 87%.

9.2 System performance

To evaluate the performance of the current version of the complete system, we have used the spoken data set collected during the usage study described above (Spechbach et al., 2019). Since the system relies on human pre-translation, it is sufficient to evaluate the output in terms of backtranslation, as a correct core-sentence will result in a correct translation for the patient. We measured the performance using sentence error rate (SER), which is defined as the percentage of core-sentences that are not identical to the annotated correct core-sentences. Since the system interface presents a selection of core-sentences to the doctor, for this evaluation we considered 3-best backtranslation results, including the GRXML result when it was above the confidence threshold and two or three backtranslations of large vocabulary recogniser results. With this configuration, the system achieved 5% SER on this data set.

10 Conclusion

Healthcare translation is required to facilitate the engagement with people with diverse language, cultural, and literacy backgrounds. The development of culturally effective and patient-oriented translation tools has become increasingly urgent. Although BabelDr is far from solving the problem of miscommunication, it is an example of a concrete application of natural language processing to help minority groups communicate in a medical context.

The developed tool, resources and evaluations are a first step toward accessible healthcare apps. This research is essential to define criteria which can be used in the development and evaluation of new medical interpreting technologies with a view to enhancing the usability among patients from refugee, migrant, or other socioeconomically disadvantaged populations.

Acknowledgements

This project was supported by the "Fondation Privée des Hôpitaux Universitaires de Genève". We would also like to thank Nuance Inc for generously making their software available to us for research purposes.

References

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

Pierrette Bouillon, Bastien David, Irene Strasly, and Hervé Spechbach. 2021. A speech translation system for medical dialogue in sign language - questionnaire on user perspective of videos and the use of avatar technology. In *Proceedings of the 3rd Swiss Conference on Barrier-free Communication (BfC 2020)*, Winterthur, Switzerland.

Pierrette Bouillon, Johanna Gerlach, Hervé Spechbach, Nikos Tsourakis, and Ismahene S. Halimi Mallem. 2017. *BabelDr vs Google Translate: a user study at Geneva University Hospitals (HUG)*, 20th Annual Conference of the European Association for Machine Translation (EAMT). Prague, Czech Republic. ID: unige:94511.

Glenn Flores, M. Barton Laws, Sandra J. Mayo, Barry Zuckerman, Milagros Abreu, Leonardo Medina, and Eric J. Hardt. 2003. Errors in medical interpretation and their potential clinical consequences in pediatric encounters. *Pediatrics*, 111(1):6–14.

Johanna Gerlach, Pierrette Bouillon, Rovena Troqe, Sonia Halimi, and Hervé Spechbach. 2021. *Patient acceptance of translation technology for medical dialogues in emergency situations*, Translation in Times of Cascading Crisis. Bloomsbury Academic.

Johanna Gerlach, Hervé Spechbach, and Pierrette Bouillon. 2018. *Creating an Online Translation Platform to Build Target Language Resources for a Medical Phraselator*, Proceedings of the 40th edition of Translating and the Computer Conference (TC40), pages 60–65. AsLing, The International Association for Advancement in Language Technology, Geneva. ID: unige:111776.

John Glauert and Ralph Elliott. 2011. Extending the sigml notation: A progress report. In *Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT)*, Dundee, Scotland.

Sonia Halimi, Razieh Azari, Pierrette Bouillon, and Hervé Spechbach. 2020. Pee or urinate? a corpus-based analysis of medical communication for context-specific responses. Corpus exploration of lexis and genres in translation. Routledge, Taylor & Francis Group.

Anissa Hamza. 2019. *La détection et la traduction automatiques de l'ellipse : enjeux théoriques et pratiques*. Ph.D. thesis, Université de Strasbourg STRASBOURG.

Antony A. Janakiram, Johanna Gerlach, Alyssa Vuadens-Lehmann, Pierrette Bouillon, and Hervé Spechbach. 2020. *User Satisfaction with a Speech-Enabled Translator in Emergency Settings*, Digital Personalized Health and Medicine, pages 1421–1422. IOS. ID: unige:139233.

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 Volume 2: Translator, Project and User Tracks*, pages 196–203, Dublin, Ireland. European Association for Machine Translation.

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, Pierrette Bouillon, Johanna Gerlach, and Hervé Spechbach. 2020. *Évaluation de la compréhension de pictogrammes Arasaac et Sclera pour améliorer l'accessibilité du système de traduction médicale BabelDr*, Handicap 2020 : technologies pour l'autonomie et l'inclusion, pages 179–182. ID: unige:144565; 11e conférence de l'IFRATH sur les technologies d'assistance.

Anita Panayiotou, Anastasia Gardner, Sue Williams, Emiliano Zucchi, Monita Mascitti-Meuter, Anita MY Goh, Emily You, Terence WH Chong, Dina Logiudice, Xiaoping Lin, Betty Haralambous, and Frances Batchelor. 2019. Language translation apps in health care settings: Expert opinion. *JMIR Mhealth Uhealth*, 7(4):e11316.

Manny Rayner, Nikos Tsourakis, and Johanna Gerlach. 2017. Lightweight spoken utterance classification with cfg, tf-idf and dynamic programming. *In: Camelin N., Estève Y., Martín-Vide C. (eds) Statistical Language and Speech Processing. SLSP 2017.*

Mark Seligman and Mike Dillinger. 2013. Automatic speech translation for healthcare: Some internet and interface aspects. In *Proceedings of 10th International Conference on Terminology and Artificial Intelligence (TIA-13)*, Paris, France.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2017. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. *CoRR*, abs/1712.05884.

Hervé Spechbach, Johanna Gerlach, Sanae Mazouri Karker, Nikos Tsourakis, Christophe Combescure, and Pierrette Bouillon. 2019. A speech-enabled fixed-phrase translator for emergency settings: Crossover study. *JMIR Med Inform*, 7(2):e13167.

Ludovic Tanguy, Cécile Fabre, Lydia-Mai Ho-Dac, and Josette Rebeyrolle. 2011. Caractérisation des échanges entre patients et médecins : approche outillée d'un corpus de consultations médicales. *Corpus, 10 |2011*, pages 137–154.

Nikos Tsourakis, Rovena Troqe, Johanna Gerlach, Pierrette Bouillon, and Hervé Spechbach. 2020. An albanian text-to-speech system for the babeldr medical speech translator. In *Digital Personalized Health and Medicine - Proceedings of MIE 2020, Medical Informatics Europe, Geneva, Switzerland, April 28 - May 1, 2020*, volume 270 of *Studies in Health Technology and Informatics*, pages 527–531. IOS Press.

Anne M Turner, Yong K Choi, Kristin Dew, Ming-Tse Tsai, Alyssa L Bosold, Shuyang Wu, Donahue Smith, and Hendrika Meischke. 2019. Evaluating the usefulness of translation technologies for emergency response communication: A scenario-based study. *JMIR Public Health Surveill*, 5(1):e11171.