Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1023–1031
Marseille, 20-25 June 2022

A Spoken Drug Prescription Dataset in French for Spoken Language Understanding

Ali Can Kocabiyikoglu[1,3] **François Portet**[1], **Prudence Gibert**[2], **Hervé Blanchon**[1], **Jean-Marc Babouchkine**[3], **Gaëtan Gavazzi**[4]
[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG,
38000 Grenoble, France
{ali-can.kocabiyikoglu, francois.portet, herve.blanchon}@univ-grenoble-alpes.fr
[2] CHU Grenoble Alpes, Avenue Maquis-du-Grésivaudan, 38700 La Tronche, France
pgibert@chu-grenoble.fr
[3] Calystene SA, 16 Rue Irène Joliot Curie, 38320 Eybens, France
jm.babouchkine@calystene.com
[4] Clinique de médecine gériatrique, CHU Grenoble Alpes, Équipe Grépi, EA 7408,
CS 10217, 38700, La Tronche, France
ggavazzi@chu-grenoble.fr

Abstract

Spoken medical dialogue systems are increasingly attracting interest to enhance access to healthcare services and improve quality and traceability of patient care. In this paper, we focus on medical drug prescriptions acquired on smartphones through spoken dialogue. Such systems would facilitate the traceability of care and would free clinicians' time. However, there is a lack of speech corpora to develop such systems since most of the related corpora are in text form and in English. To facilitate the research and development of spoken medical dialogue systems, we present, to the best of our knowledge, the first spoken medical drug prescriptions corpus, named PxSLU. It contains 4 hours of transcribed and annotated dialogues of drug prescriptions in French acquired through an experiment with 55 participants experts and non-experts in prescriptions. We also present some experiments that demonstrate the interest of this corpus for the evaluation and development of medical dialogue systems.

Keywords: Speech Corpora, Spoken Dialogue Systems, Natural Language Understanding, Health Informatics

1. Introduction

The use of information technology in healthcare has become quite prevalent in the previous years. One of the areas that has largely attracted attention is health dialogue systems used by health professionals, consumers and patients (Bickmore and Giorgino, 2006). Health institutions mostly use Hospital Information Systems (HIS) which have become essential to improve the organization and the quality of care by digitalizing nearly the entire chain of information related to the patient. One of the major components of an HIS is the Prescription Assistance Software (PAS). However, entering information to HIS is time consuming and HIS computers are sometimes far from the point of care. To deal with this situation, we propose to provide a Natural Language interface to the PAS using a smartphone. Such an interface would enable medical practitioners to enter their prescriptions orally at the point of care. Furthermore, this form of interaction would be closer to their usual practice. Such utterance would then be analyzed by Spoken Language Understanding (SLU) to send structured data to a PAS which would validate or not the prescription. Interacting through dialogue would enable the practitioner to enter her prescription quickly while leaving the system some control to make sure no legal information is forgotten (e.g. "can you specify the duration of the treatment?").

Dialogue systems, whether they are trained in an end-to-end (Zhao and Eskenazi, 2016) fashion or in a modular way (Williams et al., 2016), require large amounts of annotated data from both human-human and human-machine interactions, using natural or unnatural or constrained settings (Serban et al., 2015). Even though there is an increasing interest in building systems using publicly available datasets and improving benchmarks for general-domain dialogue systems such as in bAbI tasks (Bordes et al., 2016), the distribution of datasets in the biomedical domain is quite limited. For example, the methodical review of (Wu et al., 2020) shows that NLP related biomedical research involves a lot of private datasets which are rarely shared or replicated due to patient privacy concerns. This situation is even more difficult for languages other than English which can be subject to different regulations.

This paper presents PxSLU Corpus[1] , a drug prescription dataset comprising around $4h$ of speech recordings acquired from human-machine interactions using a prototype of a goal-oriented dialogue system used for prescribing medicine. The experiment has been performed in wild conditions with naive participants and medical experts. In total, the dataset includes 1981 recordings of 55 participants (38% non-experts, 25% doctors, 36% medical practitioners), manually tran-

[1]Our dataset is available at: `https://doi.org/10.5281/zenodo.6482586`

scribed and semantically annotated. The corpus is made publicly available through a Attribution 4.0 International (CC BY-4.0) license. It is distributed in an aligned format ready for developing and evaluating Spoken Language Understanding (SLU) systems (*conll* format). In this paper, we describe the spoken drug prescription task using smartphones, the data collection protocol and the analysis of the collected data. Furthermore, we present some Natural Language Understanding (NLU) experiments with this acquired data using recent NLP models to show the interest of such dataset for developing NLP technology for health. To the best of our knowledge, the presented dataset is the first corpus of spoken medical prescriptions fully annotated to be distributed to the community.

Outline. This paper is organized as following: Section 2 presents the related corpora while Section 3 introduces the spoken dialogue system which enables medical practitioners to record medical prescriptions through a smartphone. Section 4 explains the data acquisition protocol using our prototype dialogue system. The results of the data collection and their annotation are presented Section 5 together with experiments comparing recent NLU models trained on external data and evaluated on PxSLU. Finally, Section 6 concludes this work and gives some perspectives.

2. Medical Corpora Related to Prescriptions for NLP

2.1. Shared Tasks for Challenges in NLP for Clinical Data

In the biomedical NLP domain, there are mainly two types of public datasets: first type is the big institutional data warehouses and the second type is the datasets that are collectively built during biomedical challenges and academic datasets for a specific task.

Most of the biomedical NLP research uses big institutional warehouse datasets such as MIMIC-III (Johnson et al., 2016) or AP-HP Health Data Warehouse[2]. These datasets are distributed as a database with deidentified information and are used for numerous tasks.

On the other hand, there has been a considerable effort in creating challenges for specific tasks. Such challenges involve and stimulate data collection, annotation, evaluation and tools that are open to the scientific community. The most commonly used challenge datasets are I2B2 (Informatics for Integrating Biology and the Bedside), N2C2 (National NLP Clinical Challenges), and SemEval. Even though most of these datasets are in English, similar challenges exists in other languages than English such as QUAERO corpus (Névéol et al., 2014) for Named Entity Recognition (NER) in French. The datasets contain texts that are more substantial than simple prescriptions, such as discharge summaries or Electronic Health Records (EHRs).

[2] https://eds.aphp.fr/

2.2. Drug Prescription Datasets

Regarding drug prescriptions, we have searched for datasets that could include either whole prescriptions (preferably speech data) and that would contain prescription information available in free text. Although not many, some datasets include drug prescriptions written in natural language. For example, (Tao et al., 2018) proposes a semi-supervised prescription extraction system based on information extraction data from medical reports (Uzuner et al., 2010b). There are other challenges that include medical prescriptions mostly in narrative form inside EHRs such as I2B2 medication extraction challenge (Uzuner et al., 2010a) and Medication and Adverse Drug Events Challenge 2019 (Jagannatha et al., 2019). Another source of prescriptions could be MedDialog dataset (Chen et al., 2020). It is composed of 0.26 million medical consultations in English and 1.1 million in Chinese scrapted from online platforms. Each dialogue contains a description of patient's medical condition, a conversation between a patient and a physician and optionally diagnosis and treatment suggestions. However, the prescription part is not annotated and the dataset is only textual. To the best of our knowledge, there is no dataset of spoken drug prescriptions expressed in a natural way in any language. Therefore we aim to provide to the community a corpus with speech recordings, textual alignments and semantic annotations.

3. Understanding Spoken Medical Prescriptions

The Spoken Medical Prescriptions Understanding task follows the work of (Kocabiyikoglu et al., 2019). In this work the Spoken Medical Prescriptions are performed through a cooperative dialogue between a human prescriber and a dialogue system on a smartphone. The starting point of the dialogue is the user who initiates a dialogue session in order to record a drug prescription. The example dialogue shown Figure 1 illustrates the steps of the dialogue from the initiation to the validation.

In step **1**, the system extracts the semantics of the user's utterance by (*slot-filling*). For each attribute, the system extracts a slot-label, slot-value and a value. For example, the form of the drug destined for the patient is denoted as: *slot-filling* = (dos_uf,tablet,"tablet"). For more details about the semantic definition of slots, the reader is referred to (Kocabiyikoglu et al., 2019).

Generally, prescribers' queries are at the same time concise and specific. However, for a system to identify surely a drug, we need to interact with a medical drug database in order to retrieve structured information. For drugs, we distinguish the commercial name from its international non-proprietary name (INN) which is in this case alprazolam. Since a drug can have several entries in a drug database (for example, the same drug can exist as a capsule or a tablet, with different dosages, etc.) it is essential to accumulate all the information about

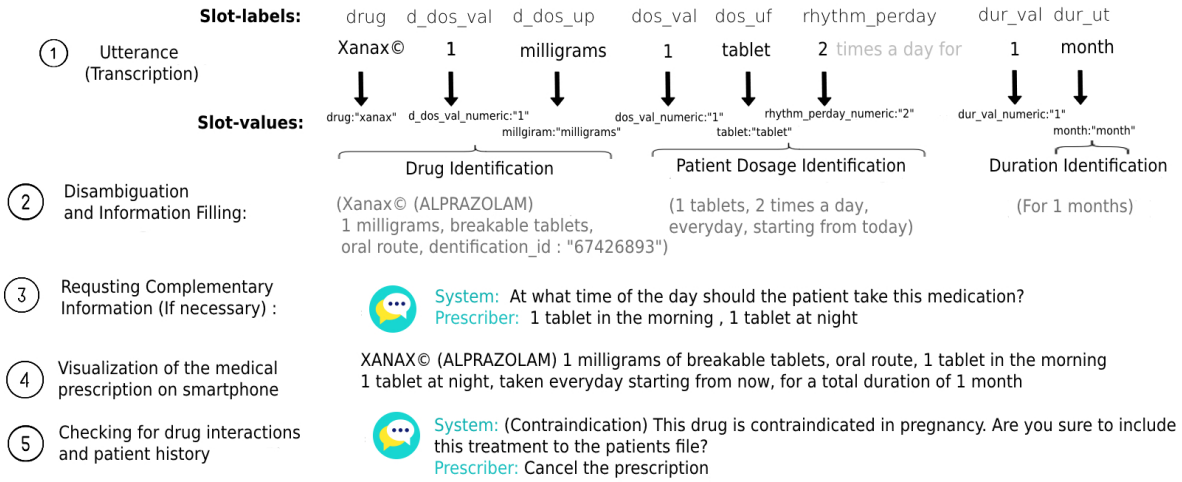

Figure 1: The steps of a dialogue between a prescriber and the dialogue system.

the drug in order to disambiguate and allow its association with a single entry. The step **2** illustrates this step which allows completing the information in a formal way (Xanax ©1 mg, breakable tablet . . .). If there are several drugs corresponding to the attributes of the current state of the dialog, the system returns a list of drugs proposed to the prescriber. If the system identifies a drug in the users' request, the dialogue continues in order to complete crucial information about the prescription.

At step **3**, the system asks the user to complete the dosage (with a valid rhythm and a frequency) and the duration of the prescription. The rhythm of the prescription defines times of the day (morning, noon, etc.) regarding drugs administration whereas the frequency defines an interval in a week (2 times a week). In our semantic definition, most of the information on the prescriptions is called non-mandatory information. For example, the fact that a drug should be taken on empty stomach is a good example of a meaningful piece of information for a prescription, but it is optional. In this step, the dialogue management module identifies the mandatory information and associates it with the appropriate dialogue act that will allow it to be collected. Once the dosage, duration and drug information are acquired, the system performs consistency checks on the prescription. In this example, "twice a day" is not specific enough, but is an anchor for the dialogue to confirm a valid rhythm and frequency for the prescription.

The step **4** concerns the validation of the prescription: the complete prescription will be shown on the screen in a textual from for explicit validation by the practitioner. Once validated, the prescription is sent to the PAS for checking.

One of the major functionalities of a PAS is its functionality to check for drug interactions and other information related to the drug and according to the patient file. In step **5**, after adding the prescription, a PAS is able to send back an alert to prescribers specifying the reason for the alert. This step is the last major step in our dialogue system. If the PAS does not find any contraindications, the prescription is added to the patient's record. The prescriber can also validate the prescription to add it to the file after having taken note of any contraindications by validating the return of PAS. In the data collection protocol, the interaction is limited to the validation of the prescription by the user and does not send the data to a PAS. Hence, step **5** was not used during the data collection experiment.

4. Data Collection Protocol

The data collection protocol that we design had an objective to collect speech data through human-machine interaction in order to train NLU models and to distribute this data within the community to allow future research and industrial development on this important application.

To allow participants to perform the data collection in their own environment, we deployed a dedicated server in the form of an API that allowed remote participation and data retrieval. This strategy enabled to collect data in a much more ecological way than inviting participants in the lab in order to record interactions in a dedicated room with an experimenter. This strategy also enabled us to perform experiments without breaching the sanitary protocol during the COVID-19 pandemic since participants could stay in their own environment using their own smartphone.

This required a lot of development and preparation as the participants had to be completely autonomous with their own smartphones. Our goal was to reach about 30 naive users and about 30 experts in drug prescriptions, including some physicians. We have established a simple protocol, inviting the participant to follow the following steps :

1. Registration on the form for requesting to participate in the experiment

2. Reception of the .apk (installation file of the mobile application on Android) and follow the document explaining the installation and the course of the experiment

3. Reception of prescription examples (depending on the audience: 20 reading examples for naive users; 10 pictographs and 10 reading examples for medical experts).

4. Filling the metadata survey (without identifying information), agree to the terms of use, and complete the experiment

Thanks to a fruitful collaboration with the University Hospital of Grenoble, we were able to involve physicians and pharmacists despite the pandemic. Non-expert participants were native French who were either members of the lab of within the social network of the co-authors.

Figure 2 shows the mobile dialogue interface and the demographic information we required. The General Terms and Conditions of use were similar to that of Common Voice[3], a large-scale effort of the Mozilla Foundation aiming at collect speech from various languages of the planet in order to make it available through a Creative Commons 0 license. The whole experiment has been discussed with the data protection officer of the Grenoble Alpes University and is registered and conforms to the European General Data Protection Regulation (GDPR).

4.1. Data Preparation

Since we targeted two types of participants – medical practitioners (doctors, pharmacists, biomedical engineers, nurses, etc.) and non-experts– all the participants did not have same expertise on medical prescriptions. Hence we did not provide the same stimuli for these two categories.

For non-expert users, we prepared a reading exercise that did not require any domain knowledge. However, for experts we wanted to be as close as possible to their own verbalization. Hence, we defined a method using iconic representations. Indeed, in order to prevent influencing the expert's utterance, we provided representations of drug prescriptions in the form of diagrams that approximates drug taking timetables. These timetables are designed for patients, generally those who are taking multiple medications, to remind the dosage and times for each medication associated with some conditions and constraints. Figure 3 shows an example of this representation. Such graphical representation allows limiting linguistic priors during the experiment.

In Figure 3, the drug (Modopar ©) is explicitly given in written text. However, in order not to influence prescribers with the days of the week, the days are represented as (D1,D2,D3,D4,D5,D6 and D7). The dosage

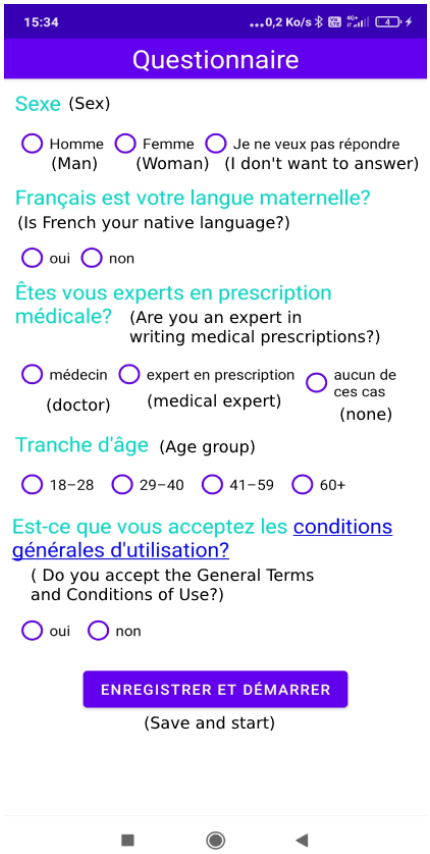

Figure 2: Screenshot of the metadata form of the main interface screen with its English translation (The terms and conditions were accessible by clicking on the link)

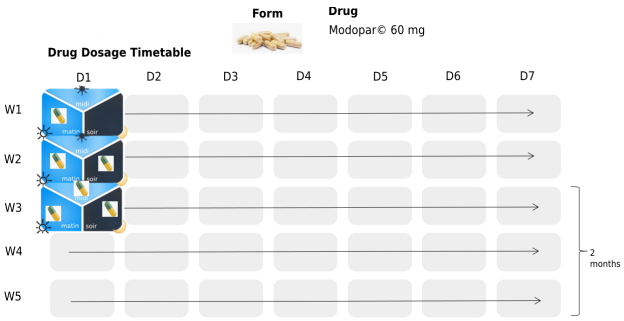

Figure 3: Example of a pictograph representation of a medical prescription

form of the drug is represented with an image, which in our example represents capsules but nothing prevents a participant to refer to it as pills or tablets. The names of the drugs, the conditions or the administration details are given in text form at the top of the screen for the prescriber to incorporate into their prescription. The dosage is indicated in the form of a calendar with boxes indicating the start time and their continuity in time. The dosage indicated in the example below denotes the progressive taking of one capsule of the drug in the morning for 1 week, then one capsule in the morn-

[3] https://commonvoice.mozilla.org/fr/terms

ing and evening for another week, then one capsule in the morning, noon and evening for 2 months. Even though this representation allows prescribers to record prescriptions that are more natural, it has the disadvantage of taking more time than a reading exercise. That's why, after 10 pictographs, the experts are asked to perform 10 reading exercises. The non-experts are directly provided with 20 textual stimuli to read.

To prepare the stimuli, real examples of prescriptions were extracted from books (especially therapeutic books) destined for students in medicine such as (Schlienger, 2013), (Delcroix and Gomez, 2020), (Denis Vital, 2018), (André, 2019), (Delcroix and Gomez, 2020) and discussed with our experts (two of which are co-authors of this paper). When prescriptions were not complete, we added duration, rhythm and frequency information with plausible values. Given the number of participants targeted, the material generated for the experiment represented approximately 300 examples of pictograph and 1300 textual drug prescriptions. Our preparation included ranking of the prescription according to their complexity (i.e., the longest and the ones with several dosage changes were last).

Based on experience from a previous study (Kocabiyikoglu et al., 2020), the total duration of the experiment was estimated, from setup to data transmission and finalization of the experiment, to 30 minutes.

4.2. Recordings Using the Spoken Dialogue System

The experiment begins with the metadata survey explained in Section 4. The metadata is saved in a local *sqlite* database in the cache directory of the mobile application. To complete the prescription entry, participants use the "Push-to-Talk" button mechanism that triggers the audio recording, and then again to stop the recording.

Figure 4 shows an example of a dialogue session recorded with the mobile application. The dialogue session starts with the participant's utterance. The speech local recording is sent to our dedicated server via a secure connection (https). From the server, the recording is then analyzed by Google's automatic speech recognition service. When the server receives the result of the speech recognition (the transcript), it is analyzed by the dialogue system which extracts the intent, semantics and tries to associate the drug-related *slots* with the drug database and determines the continuation of the dialogue which is sent back to the participant's smartphone. The dialogue continues by requesting missing information or responding to the participant's modification requests. A dialogue is completed when the participant validates a complete prescription or cancels it. In the example shown Figure 4, the participant gives the rhythm of the prescription as "2 times a day" however the system requires a more specific time of the day. Then, at dialogue turn **(4)**, the prescriber gives a more specific time which allows for the system to go one step further and show on the screen the drug

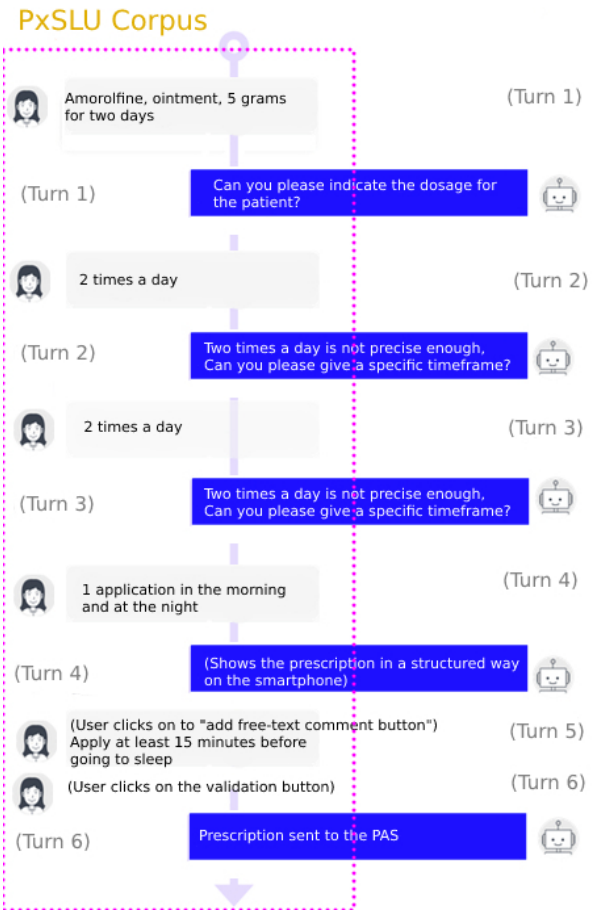

Figure 4: A dialogue example of a dialogue session

prescription in a structured way.

At the turn **(5)**, a complete prescription is shown to the participant for validation. Figure 5 shows the screen capture of the prescription. At this step, if there is a missing or incorrect information, the participant can request to correct the prescription by spoken utterance using the dialogue system. If it's a minor error, or if the participant should add additional non-structured information such as in the example above "apply at least 15 minutes before going to sleep", she can click the add a comment button (*ajouter un commentaire*) on 5 to record the message and include its transcription. Finally, the last step of a dialogue session is when the participant accepts or refuses a prescription by clicking validate (*Valider*) or cancel (*Annuler*). When all of the dialogue sessions are complete, the user can click on the upload button at the top right corner of the screen to transmit the local database containing statistics and logs and to finish the experiment.

5. Results

5.1. Collected Data

The experiment was performed between January 2021 and October 2021 (10 months of experimentation). At the end, 55 databases containing 903 dialogue sessions with 1981 sound recordings were collected. Table 1

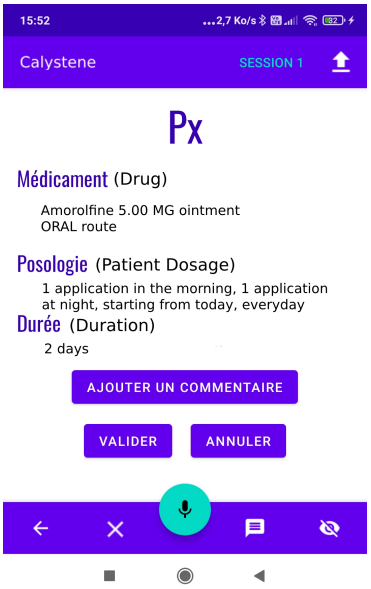

Figure 5: Visualization of a prescription on smartphone

gives a detailed overview of the collected data. The data represents 262 minutes of recordings when all the participants are included. Even though non-experts had initiated more dialogue sessions, total recording time of medical experts and doctors (∼200 minutes) are much more than non-experts (∼62 minutes). This can be explained by the fact that non-experts simply had to read the textual prescriptions and were more likely to exhaust the list than the experts who had a more time consuming interaction with the dialogue system during the phase with pictographs.

	Sessions	Recordings	Time (m)
Medical experts	258	434	94.83
Doctors	230	570	105.21
Non experts	415	977	62.13
Total	903	1981	262.27

Table 1: Overview of the collected data

The data distribution according to several participants' features is presented Figure 6. **(A)** shows that the collected dialogues are evenly distributed among the participant categories. The pie chart in **(B)** represents the distribution of the data in relation to age ranges which shows that 3 age ranges are fairly represented while the over 60+ year-old range represents only 10% of the participants. Finally, **(C)** shows that the gender representation (M/F) is well balanced.

5.2. Transcription and Semantic Labeling

The speech transcription and the semantic annotation of all the dialogues were performed by two native speakers supervised by the co-authors during the summer 2021. The two annotators were provided with Transcriber (Barras et al., 1998) and Elan (Hellwig and Van Uytvanck, 2003) tools and a document describing the transcription convention. Half a day of training was provided by the co-authors. The data consisted in the raw audio recordings and the automatic transcription that was performed by the ASR during the experiment. The task not only consisted in correcting mistakes, but also to obtain transcriptions that are closer to speech utterances. Automatic transcriptions usually remove disfluencies such as repetitions, false starts, etc. but the inclusion of these markers could be advantageous for SLU. The transcription rules made clear all the encountered cases and how to transcribe them. All transcriptions were in lower case and without punctuation. As the dialogue was obtained from human-machine interaction, the dialogue context can influence the transcription process. In order to limit this, and facilitate the transcription process, the corpus was divided into 9 batches of 100 dialogue sessions.Afterwards, these batches were divided equally between the two annotators.

Semantic annotations were performed using the doccano platform (Nakayama et al., 2018) who offers a simple graphical interface enabling to annotate data by clicking and labeling. The labeling process included 5 types of intents and 40 semantic labels that characterize drug prescriptions. Table 2 summarizes the characteristics of the NLU annotations. At the end, 14068 instances of slot-labels and 1981 instances of intents were labeled. The detailed slot-label distribution of the PxSLU corpus could be found in the appendix 7. Only 5 slots out of the 40 slot labels had fewer than 12 instances. All other slots had from 5 to 1831 instances.

Utterances	Tokens	Slots	Intents
1981	22440	14068	1981

Table 2: Summary of the semantically annotated PxSLU corpus.

5.3. Corpus Analysis

Transcription errors have a significant impact on SLU systems and the decisions of the dialogue system. For this reason, we first evaluated the word error rate (WER) performance of automatic transcription against reference transcripts. Table 3 presents the WER scores for the three categories.

	Experts	Doctors	Non experts
WER	21.99%	28.76%	24.42%

Table 3: WER scores of automatic transcriptions

The WER between the automatic and reference transcripts presented in 3 is high both for non-experts as for medical experts. The main reason for this difference is related to the labeling convention which included the transcription of onomatopoeias, disfluencies, false starts, ... whereas in the automatic transcriptions, these phenomena were discarded. Also, in the

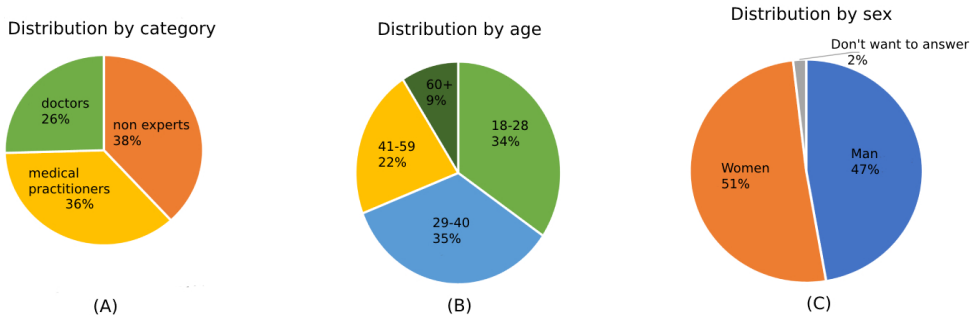

Figure 6: Distribution of some characteristics of the participants

reference transcriptions, numerical expressions regarding drug prescriptions were transcribed all in alphabetic string whereas the ASR used either numerical or alphabetic depending on the context.

In another subsequent analysis, we inspected the elapsed time at a dialogue session level. A dialogue session consists of a user starting a drug prescription until validating or refusing the prescription or restarting the session. Figure 7 shows the histogram of average elapsed time for all of the participants.

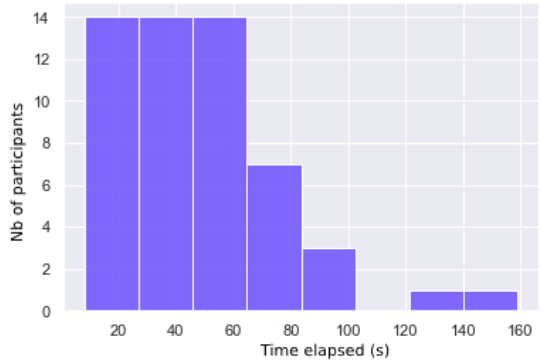

Figure 7: Histogram of average time of session by number of participants

Most of the participants completed the task in less than a minute except for a few participants who have a longer dialogue sessions that goes up to 160 seconds, which increases the average elapsed time on the metrics. Our detailed analysis on our three categories of participants show that medical practitioners have spent on average less than 40 seconds in a dialogue session which is a lot less than other categories. This might be explained by the fact that most medical practitioners that participated were pharmacists. As they have more technical knowledge about drugs, the pronunciations were easier for the system to understand. This is confirmed by the lowest WER score presented in Table 3. Furthermore, we found that more specific information about the drugs were given which resulted in reducing the time passed by the slot-filling for missing information. Thus, the validation or the refusal of a prescription was faster.

Additionally, we have noticed that doctors have spent more time on dialogue sessions (avg. 40 seconds - 100 seconds). The average number of dialogue turns per doctors is also higher than other categories. In fact, we have seen that doctors tried to interact more with the dialogue system to correct and add additional information to the visualized prescriptions which also resulted in more validated prescriptions.

5.4. NLU Model Evaluation

Our model evaluation builds on previous work (Kocabiyikoglu et al., 2019) where we have presented initial NLU systems trained on artificial and textbook data. The training data size was 35676 examples (most of them artificially generated by a context-free grammar). The table 4 shows the distribution of these examples by intent.

Intent	Examples
medical_prescription	8833
request_restart	95
negate	12608
replace	12624
none	1516

Table 4: Distribution of the training corpus

The NLU models include a classic CRF model, triangular CRF extension (Tri-CRF) (Jeong and Lee, 2008) and Bi-RNN with attention (Att-RNN) (Liu and Lane, 2016). Our recent findings on medication information extraction from EHRs in English has shown that transformer-based language models can be extremely competitive (Kocabiyikoglu et al., 2021). Even though, there are no available language models trained specifically on biomedical domain, we have decided to include Flaubert, a general-purpose pre-trained transformer language model for French (Le et al., 2019).
Table 5 shows the precision, recall and f-measure including micro, macro measures of these NLU systems on PxSLU corpus. The macro average measures are very important since it considers all the equally important *slots* whatever their frequency. Indeed, given the nature of prescriptions, most of the slots are optional and hence occur far less frequently than the mandatory ones.

The results show that the performance of the model *Flaubert* gives the best results both on micro and macro level performance. All other models have comparable performance. We can see the same behavior for the intent accuracy. However, and therefore this situation creates an imbalance problem. From the micro average perspective, the Flaubert model is by far the more robust since it performs well even for slots which are rare.

Model	Intent (acc)	Micro Avg			Macro Avg		
		P	R	F1	P	R	F1
CRF	92%	0.81	0.80	0.80	0.60	0.57	0.56
Tri-CRF	91%	0.83	0.82	0.82	0.64	0.57	0.59
Att-RNN	93%	0.83	0.87	0.85	0.55	0.55	0.53
Flaubert	**94%**	**0.89**	**0.91**	**0.90**	**0.69**	**0.74**	**0.70**

Table 5: NLU model performance on the PxSLU Corpus

Apart from providing a real test-bed to evaluate NLU model, we wanted to check if the PxSLU corpus could be used for fine-tuning the Flaubert model. For this purpose, we performed a K-Fold (K=5) cross validation with the pre-trained model "*flaubert-base-cased*". In the cross-validation process, the dataset is iteratively split into k roughly equal parts. In each iteration, each of the k part is used as a test set and the rest is used for training. In each run, *fine-tuning* is performed for three epochs. Table 5.4 shows the results of this experiment.

K #	Micro Average			Macro Average		
	P	R	F1	P	R	F1
1	0.93	0.93	0.93	**0.79**	0.75	0.75
2	0.93	0.94	0.94	0.76	**0.80**	**0.77**
3	0.89	0.90	0.90	0.54	0.48	0.50
4	0.92	0.91	0.91	0.68	0.67	0.66
5	**0.95**	**0.95**	**0.95**	0.73	0.74	0.72
avg	0.92	0.92	0.92	0.70	0.68	0.68
SD	0.02	0.02	0.02	0.10	0.12	0.10

Table 6: K-fold cross validation result of the Flaubert model on PxSLU Corpus (SD=Standard Deviation)

The results shown in 5.4 shows that a model trained on PxSLU obtain comparable results with those given in the table 5. This shows that PxSLU can be used for fine-tuning and evaluation to lead to similar performance than models training on larger artificial datasets (books as well as artificial data). Moreover, the results of the different k-folds vary more when we look at the macro average, which is confirmed by the standard deviation which is also higher. This shows that the choice of data impacts the macro performance and thus the coverage of the *slots*. A finer data partitioning could thus lead to even greater performance.

6. Conclusion

We have presented PxSLU corpus which, to the best of our knowledge, is the first drug prescription dataset of speech recordings constructed from human-machine interaction. The dataset includes about $4h$ of speech recordings collected from 55 participants with medical experts. The automatic transcriptions were verified by human effort and aligned with semantic labels to allow training of NLP models. The data acquisition protocol was reviewed by medical experts and permit free distribution without breach of privacy and regulation. The analysis of the corpus and the evaluation of recent NLU models on PxSLU showed that the dataset is realistic and can be used as a benchmark. Furthermore, it can be efficiently used to fine-tune pre-trained language models.

PxSLU can be used in many other tasks including dialogue (Kocabiyikoglu et al., 2019) and SLU. We hope that that the community will be able to benefit from PxSLU which will be distributed with a Attribution 4.0 International (CC BY 4.0) license. In a further study, we intend to present check if the dialogue obtained during the experiment can be used to evaluate and train dialogue models.

7. Acknowledgements

This work was supported by a CIFRE grant number 2017/1798 from ANRT (National Association for Research and Technology) and was partially supported by MIAI@Grenoble-Alpes (ANR-19-P3IA-0003).

Appendix: Slot-label Distribution of PxNLU Corpus

The slot-label distribution of PxSLU corpus is presented in the Table 7. As expected from our previous findings (Kocabiyikoglu et al., 2019), there is a class imbalance in the semantic information. This is because in drug prescriptions, the mandatory information (drug name, duration, dosage, etc.) are more present than additional information such as (taking on empty stomach). The slot distribution shows that there are 1285 drug prescriptions (drug+inn categories) in the recordings.

drug	1042	**inn**	431	**max-unit-ut**	22
rhythm-rec-ut	36	**rhythm-tdte**	1452	**min-gap-val**	5
d-dos-val	954	**rhythm-perday**	264	**min-gap-ut**	8
d-dos-up	924	**rhythm-hour**	119	**cma-event**	294
d-dos-form	303	**freq-val**	22	**fasting**	18
d-dos-form-ext	68	**freq-startday**	7	**rhythm-rec-val**	31
A	57	**freq-ut**	125	**re-val**	36
roa	55	**freq-days**	17	**re-ut**	30
dos-val	1775	**freq-int-v1**	31	**ns**	0
dos-uf	1679	**freq-int-v1-ut**	26	**nr**	0
dos-cond	145	**freq-int-v2**	20	**qsp-val**	30
max-unit-val	30	**freq-int-v2-ut**	10	**qsp-ut**	29
max-unit-uf	18	**dur-val**	1402	**dur-ut**	1402

Table 7: Slot-label distribution of PxSLU Corpus– Revoir la fréquence

8. Bibliographical References

André, P. (2019). *Ordonnances en parasitologie, médecine tropicale et des voyages*. Maloine.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (1998). Transcriber: a free tool for segmenting, labeling and transcribing speech. In *First international conference on language resources and evaluation (LREC)*, pages 1373–1376.

Bickmore, T. and Giorgino, T. (2006). Health dialog systems for patients and consumers. *Journal of biomedical informatics*, 39(5):556–571.

Bordes, A., Boureau, Y.-L., and Weston, J. (2016). Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*.

Chen, S., Ju, Z., Dong, X., Fang, H., Wang, S., Yang, Y., Zeng, J., Zhang, R., Zhang, R., Zhou, M., Zhu, P., and Xie, P. (2020). Meddialog: a large-scale medical dialogue dataset. *arXiv preprint arXiv:2004.03329*.

Delcroix, M.-H. and Gomez, C. (2020). *Ordonnances en gynécologie obstétrique: 103 prescriptions courantes*. Maloine.

Denis Vital, D. (2018). *Ordonnances 2019: 180 PRESCRIPTIONS COURANTES EN MEDECINE*. Maloine.

Hellwig, B. and Van Uytvanck, D. (2003). Eudico linguistic annotator (elan) version 1.4-manual. *Last updated*.

Jagannatha, A., Liu, F., Liu, W., and Yu, H. (2019). Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (made 1.0). *Drug safety*, 42(1):99–111.

Jeong, M. and Lee, G. G. (2008). Triangular-chain conditional random fields. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(7):1287–1302.

Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Kocabiyikoglu, A. C., Portet, F., Blanchon, H., and Babouchkine, J.-M. (2019). Towards spoken medical prescription understanding. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–8. IEEE.

Kocabiyikoglu, A. C., Portet, F., Babouchkine, J.-M., and Blanchon, H. (2020). Spoken medical prescription acquisition through a dialogue system on smartphone: Perspective of a healthcare software company. In *LREC 2020 Industry Track Language Resources and Evaluation Conference 11–16 May 2020*.

Kocabiyikoglu, A. C., Babouchkine, J.-M., Portet, F., and Qader, R. (2021). Neural medication extraction: A comparison of recent models in supervised and semi-supervised learning settings. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 148–152. IEEE.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2019). Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*.

Liu, B. and Lane, I. (2016). Attention-based recurrent neural network models for joint intent detection and slot filling. In *Interspeech 2016*, pages 685–689.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Névéol, A., Grouin, C., Leixa, J., Rosset, S., and Zweigenbaum, P. (2014). The QUAERO French medical corpus: A ressource for medical entity recognition and normalization. In *Proc of BioTextMining Work*, pages 24–30.

Schlienger, J.-L. (2013). *100 situations clés en médecine générale: Évaluation, Diagnostic, Thérapeutique*. Elsevier Health Sciences.

Serban, I. V., Lowe, R., Henderson, P., Charlin, L., and Pineau, J. (2015). A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.

Tao, C., Filannino, M., and Uzuner, Ö. (2018). Fable: A semi-supervised prescription information extraction system. In *AMIA Annual Symposium proceedings*, volume 2018, page 1534. American Medical Informatics Association.

Uzuner, Ö., Solti, I., and Cadag, E. (2010a). Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Uzuner, Ö., Solti, I., Xia, F., and Cadag, E. (2010b). Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, 17(5):519–523.

Williams, J., Raux, A., and Henderson, M. (2016). The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., Soni, S., Wang, Q., Wei, Q., Xiang, Y., et al. (2020). Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470.

Zhao, T. and Eskenazi, M. (2016). Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. *arXiv preprint arXiv:1606.02560*.