Natural Language Engineering **1** (1): 1–**??**. Printed in the United Kingdom 1
© 201? Cambridge University Press

Designing a Virtual Patient Dialogue System Based on Terminology-rich Resources: Challenges and Evaluation

LEONARDO CAMPILLOS-LLANOS[1,2], CATHERINE THOMAS[1,2], ÉRIC BILINSKI[1], PIERRE ZWEIGENBAUM[1], SOPHIE ROSSET[1]

[1] *LIMSI, CNRS, Université Paris-Saclay, Orsay, France,*
[2] *SATT Paris-Saclay, Orsay, France*
{*campillos,thomas,bilinski,pz,rosset*}*@limsi.fr*

(*Received* ... ; *revised* ...)

Abstract

Virtual patient software allows health professionals to practice their skills by interacting with tools simulating clinical scenarios. A natural language dialogue system can provide natural interaction for medical history taking. However, the large number of concepts and terms in the medical domain makes the creation of such a system a demanding task.

We designed a dialogue system that stands out from current research by its ability to handle a wide variety of medical specialties and clinical cases. To address the task, we designed a patient record model, a knowledge model for the task, and a termino-ontological model that hosts structured thesauri with linguistic, terminological and ontological knowledge. We used a frame- and rule-based approach and terminology-rich resources to handle the medical dialogue. This work focuses on the termino-ontological model, the challenges involved and how the system manages resources for the French language.

We adopted a comprehensive approach to collect terms and ontological knowledge, and dictionaries of affixes, synonyms and derivational variants. Resources include domain lists containing over 161,000 terms, and dictionaries with over 959,000 word/concept entries.

We assessed our approach by having 71 participants (39 medical doctors and 32 non-medical evaluators) interact with the system and use 35 cases from 18 specialities. We conducted a quantitative evaluation of all components by analysing interaction logs (11,834 turns). Natural language understanding achieved an F-measure of 95.8 per cent. Dialogue management provided on average 74.3 (±9.5) per cent of correct answers. We performed a qualitative evaluation by collecting 171 five-point Likert scale questionnaires. All evaluated aspects obtained mean scores above the Likert mid-scale point. We analysed the vocabulary coverage with regard to unseen cases: the system covered 97.8 per cent of their terms.

Evaluations showed that the system achieved high vocabulary coverage on unseen cases and was assessed as relevant for the task.

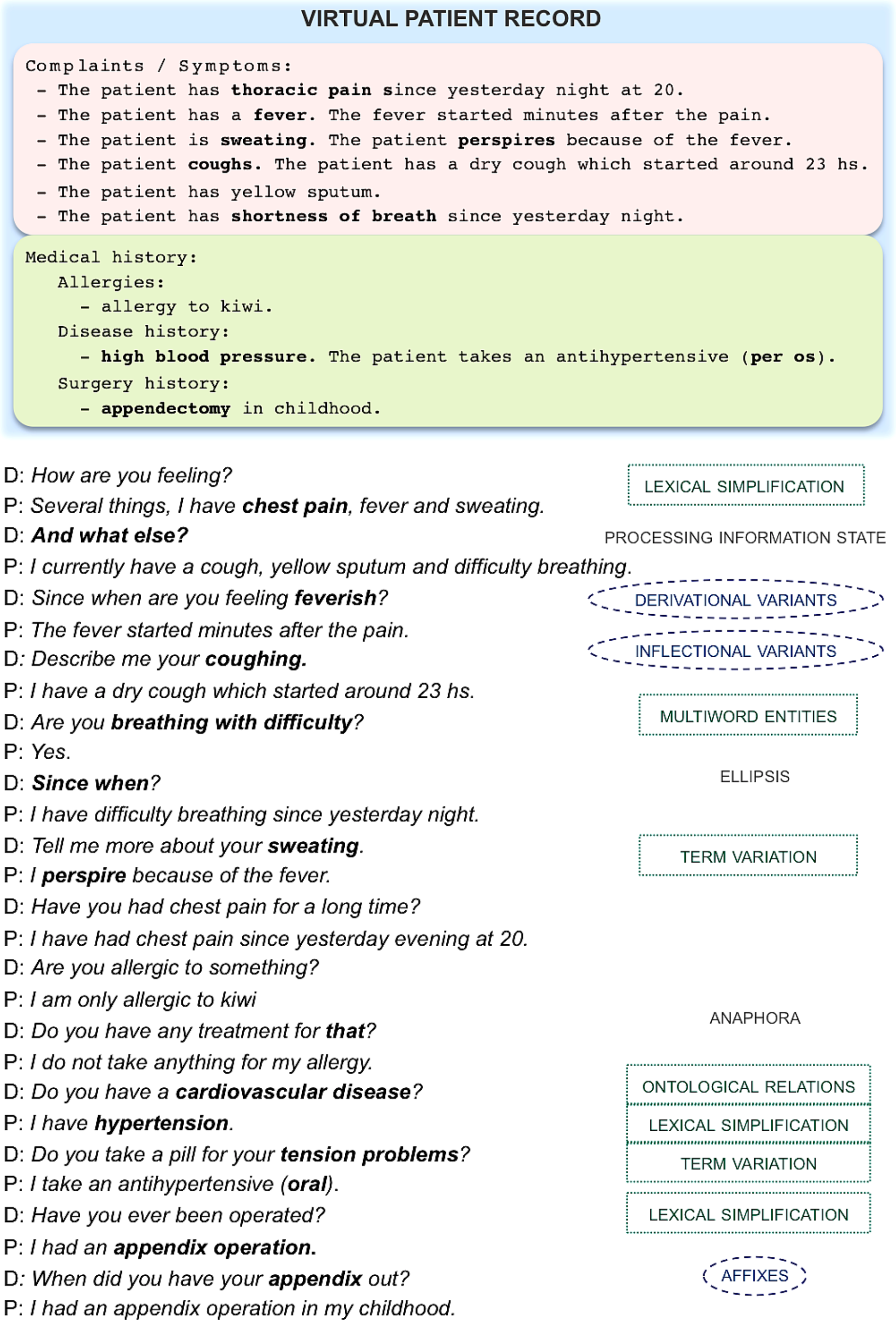
1 Introduction

Medical education requires trainees and practising doctors to develop expertise in diagnosis or clinical reasoning. These skills are traditionally acquired through clinical practice, and they may be enhanced with the help of simulations with mannequins, role-playing games or virtual patients (Rombauts 2014). More broadly, the literature uses the term *virtual patient* to refer to simulations such as case presentations, interactive patient scenarios, high-fidelity mannequins, virtual patient games, high-fidelity software simulations, human standardised patients—who are actors playing the role of interviewed patients paid for educational purposes—or virtual standardised patients (Talbot *et al.* 2012a). Virtual patients allow health professionals to practice their skills by interacting with a software 'that simulates real-life scenarios' (Cook, Erwin and Triola 2010). In our work, *virtual patient* (hereafter, VP) refers to virtual standardised patients. For the last few decades, VPs have allowed doctors to train clinical and history taking skills through simulated scenarios in digital environments (Ellaway *et al.* 2006; Danforth *et al.* 2009).

Interactivity with a VP might be enhanced through a dialogue system, but such a component needs to address several phenomena to achieve a natural, user-friendly dialogue (Figure 1). As shown, medical doctors tend to begin by eliciting initial clues from the patient by using broad questions. Then, they use follow-up questions to focus on specific details. The system needs to deal with this behaviour by processing context information (ellipsis and anaphora) and updating its information state, so that it avoids providing redundant answers. In addition, term variants referring to the same concept need to be mapped accurately (e.g. *hypertension* ↔ *high blood pressure*) by means of linguistic and terminological knowledge.

This work describes our endeavour to create a dialogue system featuring unconstrained natural language interaction in a simulated consultation with a VP. We built this system to simulate history-taking in an educational software featuring an animated avatar with text-to-speech (Figure 2), and allowing students to simulate a physical exam. This project was developed in collaboration with a medical team specialized in simulation-based medical education at Angers University Hospital (CHU d'Angers) and several companies (Interaction Healthcare / SimforHealth, Voxygen and VIDAL®). Within this partnership,[1] each group deployed their know-how: 1) to provide the system with advanced dialogue capabilities when users interview each VP: this was our contribution, and is the background for the present paper; 2) to endow each VP with text-to-speech and minor gestures; 3) to develop a dedicated interface for instructors to create unlimited avatars and patient records; 4) to manage the evaluation with potential end-users who tested the system; and 5) to provide domain texts and data for the task. In this project, which spanned over three years, four computational linguists and one software engineer (about 4.5 person-years) were involved to create the dialogue system.

[1] More information about the project is available at: `https://pvdial.limsi.fr/`



VIRTUAL PATIENT RECORD

Complaints / Symptoms:
- The patient has **thoracic pain** since yesterday night at 20.
- The patient has a **fever.** The fever started minutes after the pain.
- The patient is **sweating.** The patient **perspires** because of the fever.
- The patient **coughs.** The patient has a dry cough which started around 23 hs.
- The patient has yellow sputum.
- The patient has **shortness of breath** since yesterday night.

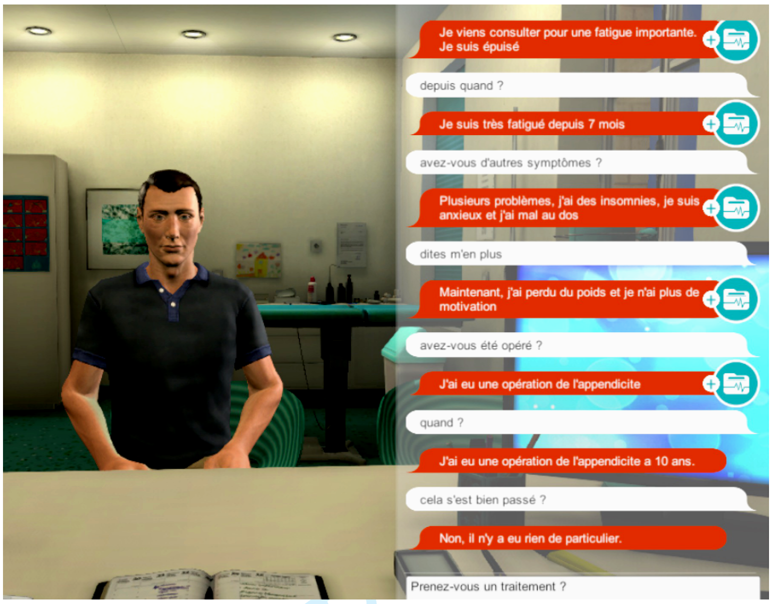
Medical history:
Allergies:
- allergy to kiwi.

Disease history:
- **high blood pressure.** The patient takes an antihypertensive (**per os**).

Surgery history:
- **appendectomy** in childhood.

D: *How are you feeling?*
P: *Several things, I have **chest pain**, fever and sweating.*
D: ***And what else?***
P: *I currently have a cough, yellow sputum and difficulty breathing.*
D: *Since when are you feeling **feverish**?*
P: *The fever started minutes after the pain.*
D: *Describe me your **coughing.***
P: *I have a dry cough which started around 23 hs.*
D: *Are you **breathing with difficulty**?*
P: *Yes.*
D: ***Since when**?*
P: *I have difficulty breathing since yesterday night.*
D: *Tell me more about your **sweating**.*
P: *I **perspire** because of the fever.*
D: *Have you had chest pain for a long time?*
P: *I have had chest pain since yesterday evening at 20.*
D: *Are you allergic to something?*
P: *I am only allergic to kiwi*
D: *Do you have any treatment for **that**?*
P: *I do not take anything for my allergy.*
D: *Do you have a **cardiovascular disease**?*
P: *I have **hypertension**.*
D: *Do you take a pill for your **tension problems**?*
P: *I take an antihypertensive (**oral**).*
D: *Have you ever been operated?*
P: *I had an **appendix operation.***
D: *When did you have your **appendix** out?*
P: *I had an appendix operation in my childhood.*

Fig. 1: Sample dialogue (D: doctor; P: patient, simulated by the system) and relation to the patient record. Phenomena to be addressed are shown in bold: discourse phenomena, linguistic variation (blue circles) and termino-ontological variation (green boxes). The patient record structure is simplified due to space constraints. We show real replies of the current version of the French system, but we translated them to English.



Je viens consulter pour une fatigue importante. Je suis épuisé

depuis quand ?

Je suis très fatigué depuis 7 mois

avez-vous d'autres symptômes ?

Plusieurs problèmes, j'ai des insomnies, je suis anxieux et j'ai mal au dos

dites m'en plus

Maintenant, j'ai perdu du poids et je n'ai plus de motivation

avez-vous été opéré ?

J'ai eu une opération de l'appendicite

quand ?

J'ai eu une opération de l'appendicite a 10 ans.

cela s'est bien passé ?

Non, il n'y a eu rien de particulier.

Prenez-vous un traitement ?

Fig. 2: Sample of dialogue with a virtual patient avatar

In this context, the medical team set specific design constraints:

- The system should manage a dialogue for unseen cases without being extended manually. Achieving *high vocabulary coverage* allows trainees to interact with VPs in a wide range of clinical conditions. The items to which a doctor must pay attention vary widely across cases and specialities. Practising in front of a variety of VPs should therefore be a good exercise to train students in medical history taking. Certainly, practising the history-taking skills with many cases needs to be accompanied by quality scenarios carefully designed by medical instructors to provide feedback and meaningful learning opportunities.
- The VP should provide answers based on the actual content of the patient record, i.e. the information representing the patient's health or findings (e.g. disease history or medication, § 3.1). This brings up the need for providing correct replies, an aspect we refer to as *correctness*. By this, we mean that the simulated patient should not invent information that is not present in the patient record (*faithfulness*). Related to that aspect, the system should not omit data from the record (*exhaustiveness of information*). In our project, it was deemed that the system should provide all data available with regard to a given question. This design criterion seemed more adequate to avoid adding more difficulties to other possible sources of miscommunication arising with an artificial agent. Note that this was a pedagogical choice for a first version of the system: a virtual patient could indeed simulate situations in real life where the patient lies or makes mistakes in recalling the health record. Medical students should keep that in mind when conducting a medical history.

We can add the following observations related to the task and domain:

- The conversation does not take place in an open context but in a specific domain. The medical-history taking is a focused task, defined by the number and nature of the topics that are to be processed (§ 3.2).
- The medical domain involves a large concept space with a multitude of term variants. This requires *entity linking* (also called *entity normalisation*) in order to query and match input terms with those in the patient record. Our goal is to develop a system that, first, replies correctly (*correctness*), and second, can work with an unlimited number of cases and medical specialities (*high vocabulary coverage*). A method to evaluate these aspects at the natural language understanding step is to compute *precision* and *recall*: precision is related to correctness, whereas recall depends on vocabulary coverage. When evaluating dialogue management, *correctness* is a key aspect.
- A lexical simplification process is needed to simulate natural replies, i.e. using terms according to a patient viewpoint (e.g. saying *appendix operation* instead of the professional term *appendectomy* written in the patient record).
- Pre-existing dialogue data for the task and domain are not available, hence the system cannot be designed with a data-intensive approach. In this sense, another contribution of our work is making available for the community a corpus of dialogue data collected during this project.[2]

Herein we explain how we address these requirements in our dialogue system and the challenges involved. To handle the needs of the domain and the diversity of its concepts, we applied a comprehensive approach to terminology collection. To the best of our knowledge, we integrated a larger volume of resources than in standard task-oriented conversational agents (see §3). We propose a framework for managing the linguistic and terminological needs in such a task, with different levels of knowledge representation, keeping coherence across components. The difficulty of term detection in our task motivates our approach based on rich terminological resources, which complement a frame- and rule-based dialogue management. The proposed methods aim to enhance the system's capability to adapt to new cases in a wide range of medical specialities and detect rare and unseen vocabulary items for a successful interaction. To show the extent to which it succeeds in doing so in a real-use scenario, we conducted evaluations which include user interactions (n=71) with 35 different VP cases from 18 different medical specialities, and an assessment of vocabulary coverage on 169 new cases.

We developed French, English and Spanish versions of the system. We report here work related to the French version because it is the only one evaluated to date. In the remainder of the paper, we first review the approaches to dialogue systems and to interactions with simulated VPs (§2). We present our termino-ontological and linguistic models in §3 and summarise the architecture of our system in §4. We present the evaluation of resources and discuss its results in §5, then conclude in §6.

[2] https://pvdial.limsi.fr/data/PG-logs-eval.zip

2 Related work

2.1 Approaches to dialogue systems

A textual or spoken dialogue system involves several modules, which typically include natural language understanding, speech recognition, generation, speech synthesis and a dialogue manager. More components may be added to correct input errors, query a database or search in a document collection, if the task requires it.

Different approaches exist for dialogue management (Jokinen and McTear 2009): finite-state or graph-based approaches (Cole 1999); frame-based techniques (McTear *et al.* 2005; van Schooten *et al.* 2007); statistically-based approaches such as reinforcement learning (Sutton and Barto 1998) or Partially Observable Markov Decision Processes (POMDP) (Young 2006); and neural-network-based approaches, which have recently been reviewed (Celikyilmaz, Deng and Hakkani-Tur 2017).

Due to the lack of existing dialogue corpora for this domain, statistically or neural approaches are not applicable to the design of our VP dialogue system. Our approach aims at endowing the system with capabilities to manage a comprehensive range of aspects of a dialogue task. This involves processing input so that it is understood in the context of previous questions and answers: interpreting discourse phenomena (such as co-reference or ellipsis) and maintaining the global dialogue state. This contrasts with question-answering approaches, designed to answer independent questions, in which a system analyses the user's natural language questions and produces a natural language answer, but commonly without processing of dialogue history (Talbot *et al.* 2016; Maicher *et al.* 2017; Jin *et al.* 2017).

2.2 User interaction in healthcare applications

Interactive systems for healthcare applications address patient education and counselling (Giorgino *et al.* 2005; Bickmore 2015) or support to practitioners (Beveridge and Fox 2006); a review is reported in (Bickmore and Giorgino 2006). We focus on a dialogue with a VP in an educational context. Development challenges are similar across health dialogue systems (Hoxha and Weng 2016). Literature reviews on VPs are available (Cook, Erwin and Triola 2010; Kenny and Parsons 2011; Salazar *et al.* 2012; Rossen and Lok 2012; Lelardeux *et al.* 2013; Rombauts 2014).

A key strategy for enhancing the simulation is to provide realistic user interaction. Integrating natural language interaction into a VP system requires managing domain terms—e.g. by formalizing ontological concepts (Nirenburg *et al.* 2008a)—and Natural Language Understanding (NLU). The NLU unit may rely on text meaning representations for resolving paraphrases (Nirenburg *et al.* 2009), a corpus of questions and replies curated by experts (Kenny *et al.* 2008) or canned questions and answers (Benedict 2010; Siregard, Julen and Lessard 2013). The i-Human Patients® system[3] allows users to choose the questions to ask to the VP, whose answers are parametrized by a patient record.

[3] http://www.i-human.com/

Few systems allow natural language input. As far as we can tell, current tools with natural language interaction are available for practicing patient assessment and diagnosis (Hubal *et al.* 2000)—e.g. in a pediatric scenario (Hubal *et al.* 2003)—and clinical history taking and communication skills—e.g. in a case of acute abdominal pain (Stevens *et al.* 2006); in a psychiatric consultation (Kenny *et al.* 2008); or in a case of back pain (Gokcen *et al.* 2016; Maicher *et al.* 2017).[4] Kenny and his team reported using 459 question variants mapped to 116 responses related to a post-traumatic stress disorder case (Kenny *et al.* 2008). Gokcen and colleagues' system partially relies on manually annotated data—to date, 104 dialogues and 5,347 total turns (Gokcen *et al.* 2016). The Maryland Virtual Patient, which simulates seven types of esophageal diseases, uses a lexicon covering over 30,000 word senses and an ontology of more than 9,000 concepts (Nirenburg *et al.* 2008b).

These interactive systems seem to be case-specific; i.e. they treat a limited number of cases. As far as we know, Talbot *et al.* (2016) developed one of the few natural language interaction systems trained to cope with different clinical cases in the English language (e.g. ear pain, psychiatry and gastroenterology).[5] It relies on a medical taxonomy of 700 questions and statements and a supervised machine-learning model trained on over 10,000 training examples. For their part, the Virtual Patients Group (VPG, a consortium of North-American universities) also envisages a robust natural language interaction system. The VPG's platform Virtual Patient Factory[6] allows users to create new cases and interact with virtual humans. Application scenarios range from psychiatry to pharmacy. To develop the NLU component, they used the Human-Centered Distributed Conversational Modeling (HDCM) technique (Rossen, Lind and Lok 2009), a crowd-sourcing methodology for collecting the corpus used to feed the system. Their method relies on a tight collaboration between VP developers and medical experts, a workflow that we specifically aim to bypass to make the system much more easily extensible to new cases.

Lastly, neural approaches are being explored for the NLU component in VP systems (Datta *et al.* 2016; Jin *et al.* 2017). These are data-intensive methods and can be set up once enough data are collected from real interactions.

We applied a knowledge-based approach, mostly rule and frame-based, because of the lack of available dialogue and domain data to train a machine learning system. Due to the magnitude of the terminology in the medical domain, we also rely on rich terminological resources, which led us to give special care to the design of language resources management.

3 Models

Given a clinical case, the medical trainee will ask questions about various facets of the patient record, referring to entity types and concepts through domain terms. In this section, we present the models designed to create our VP dialogue system.

[4] The OSU VP Project: `http://128.146.170.201/WEBGL/JackWilson/`
[5] The system can be tested at: `https://prod.standardpatient.org/`
[6] `http://www.virtualpeoplefactory.com/Classic/Home`

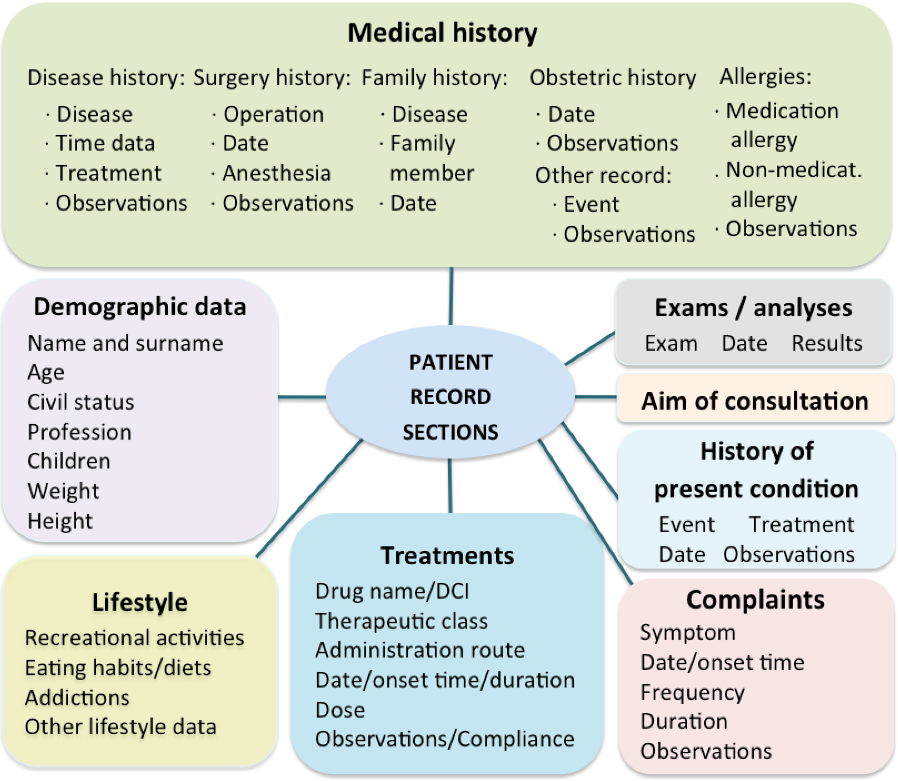


Fig. 3: Patient record model

First, the patient record model (§3.1) defines the patient's health status. Second, the knowledge model of the dialogue task (§3.2) defines the range of questions types, entity types and dialogue acts. Third, the linguistic and termino-ontological model (§3.3) defines domain relations and concepts and manages linguistic variation.

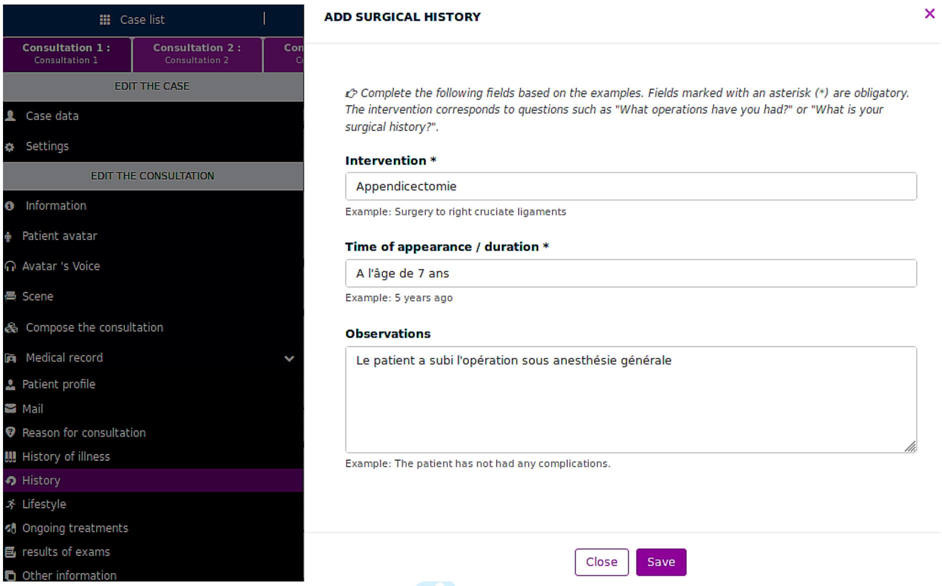
3.1 Patient record model

We first need to specify what type of information is available to the dialogue system about the patient's state. This is the aim of the *patient record model*, from which the dialogue system will provide information about the specific patient it embodies. This model is similar to those that underlie electronic health records, and was refined iteratively when the first cases were created.

The VP is specified in a clinical record typically authored by a medical trainer, who aims at teaching students how to handle a given case. The clinical record describes common data found in patient records, structured into sections (e.g. Medical history or Current treatment) and lower-level subsections (e.g. Medical history has a subsection on Family history; Figure 3). Some basic elements can have associated attributes (e.g. the dose and frequency of a medical treatment). Most elements and attributes consist of free text. An example of a clinical record is shown in Table 1.

```
aimOfConsultation:
    aim: the patient felt a sudden, intense pain on the right side
informations:
    patientFirstName: Martin
    patientLastName: Tournier
    patientAge: 69
    patientSex: man
    maritalStatus: married
    profession: retired, former taxi driver
    children: 3 children
    weight: 78
    height: 1 meter 74 centimeters
lifestyle:
    food:
      items:
        - the patient has normal eating habits
    physicalActivity:
      items:
        - the patient likes gardening
    addictions:
      items:
        - the patient smokes 20 cigarettes a day since 20 years old
        - the patient drinks a glass of wine per day
medicalRecord:
    allergies:
      nonmedicationAllergy:
        - allergy: kiwi
    medicalHistories:
      - disease: high blood pressure
        onsetTime: 10 years ago
        treatment:
          therapeuticClassValue: antihypertensive
          doseValue: 20 milligrams
          frequencyValue: 1 pill every day
          methodOfAdministrationValue: per os
    surgeries:
      - operation: appendectomy
        lifeperiod: childhood
      - operation: inguinal herniorrhaphy
        anesthesia: GA
        age: at 51
    other:
      - observationsValue: the patient is up-to-date with his vaccines
      - observationsValue: the patient does not live in damp housing
complaints:
    - symptom: the patient has thoracic pain on the right side
      onsetTime: since yesterday night at 20
      observationsValue: the patient had pain below the nipple, the patient was watching TV
    - symptom: the patient has a fever
      onsetTime: the fever started minutes after the pain
      feverValue: 38.9 degrees
      observationsValue: the patient did not take a medication
    - symptom: the patient is sweating
      observationsValue: the patient perspires because of the fever
      onsetTime: since yesterday night
    - symptom: the patient coughs
      observationsValue: the patient has a dry coughing since yesterday night at 23
    - symptom: the patient has yellow sputum
      onsetTime: since today morning
    - symptom: the patient has shortness of breath
      onsetTime: since yesterday night
currentTreatment:
    - therapeuticClassValue: pain-killer
```

Table 1: Sample clinical record: the format used is YAML



ADD SURGICAL HISTORY

Complete the following fields based on the examples. Fields marked with an asterisk (*) are obligatory. The intervention corresponds to questions such as "What operations have you had?" or "What is your surgical history?".

Intervention *

Appendicectomie

Example: Surgery to right cruciate ligaments

Time of appearance / duration *

A l'âge de 7 ans

Example: 5 years ago

Observations

Le patient a subi l'opération sous anesthésie générale

Example: The patient has not had any complications.

Fig. 4: Dedicated interface to input data on the virtual patient record

Medical instructors prepared patient records through a dedicated interface (Figure 4). A preprocessing module extracts information from some text fields, e.g. time data from the symptoms (complaints) field. The system most often reuses text from the patient record to generate replies, which may make them less natural.

3.2 Knowledge model for the dialogue task

Our system operates in the context of the anamnesis stage of a consultation scenario—i.e., the medical-history taking step to collect diagnostic information from a patient. Therefore, the system must be capable of analyzing and replying to common questions (related to symptoms or treatments). The system also needs to process broader designated topics that are necessary to conduct history-taking; namely, demographic data (patient's name and surname, civil status, age, profession or family status) and questions on patient's lifestyle (recreational activities, eating habits, social life or family life). The knowledge model for the dialogue task defines:

- the range of topics that might be addressed during the clinical anamnesis, and the relation to each record section (as they were grouped in our project);
- the question types to be processed in the dialogue manager, as well as the sections in the VP record associated to these question types;
- the entity types to annotate in the user input by the natural language understanding (NLU) module (§3.3.2);
- dialogue acts, which define the function of the user input (e.g., greeting, acknowledging) analysed by the NLU module, and acts that define the reply type and content to be output by the natural language generation (NLG) module (e.g. `inform_symptom_duration`).

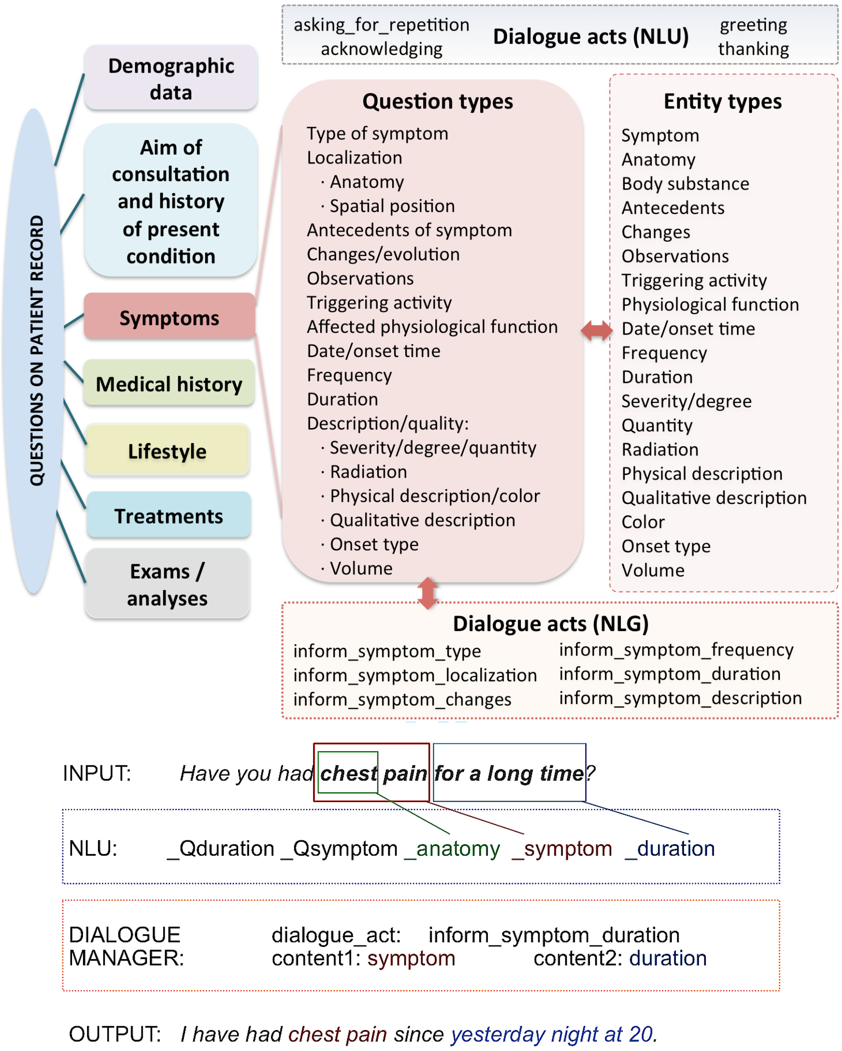


Fig. 5: Knowledge model for the dialogue task (sample). The upper part shows the relation between questions on the patient record, questions types and topics, and entity types. The lower part shows how system components instantiate this model. NLU: natural language understanding; NLG: natural language generation.

A sample of the model with regard to symptoms is described in Figure 5.

We designed the model based upon the following sources. We collected questions used in a patient-doctor consultation scenario. We used 30 audio recordings with human standardised patients, who simulated consultations on anesthesiology, hypertension and pneumopathy. Several actors were recorded for each case, which allowed us to obtain varied versions of history taking for the same case.[7] We transcribed and analysed the recordings to detect interaction patterns. We also gathered questions from guides for clinical examination used by practitioners (Bates and Bickley 2014; Epstein *et al.* 2015) and from resources for medical translation (Coudé, Coudé and Kassmann 2011; Pastore 2015).[8]

For entity types and dialogue acts, we defined 149 different labels: 62 entity types, 70 question types, and 17 dialogue acts (e.g. greetings). 52 labels are used for medical entity types and 10 for non-specifically medical entity types (e.g., `frequency`). 56 labels are used for medical question types (e.g. `Qsymptom`) and 14 for general question types (e.g. for dates, `Qdate`, *when*; or cause, `Qwhy`, *why*). Medical entity and question types are related to patient record sections: lifestyle (16.7 per cent of labels), medical history/symptoms (60.1 per cent), treatments (8.3 per cent), clinical examinations/analyses (4.6 per cent) and demographic data (10.2 per cent).

The knowledge model for the dialogue task was defined and refined in an iterative process during development: evaluators periodically interacted with the system, and a computational linguist analysed the logs to improve the questions types, the labels of entity types and the missing terms used in the interaction. The process extended over 7 iterations and about 24 months. We report the domain sources used for the lists of entity types in Table 3 (§3.3.2).

3.3 Linguistic and termino-ontological model

Medical terminology brings up multiple processing difficulties. To illustrate their order of magnitude, let us first introduce the Unified Medical Language System® (hereafter, UMLS®) (Bodenreider 2004). The UMLS MetaThesaurus® is a large multilingual source of medical terminologies and ontological knowledge that come from close to 200 thesauri. In a similar way to how WordNet encodes synsets, the UMLS MetaThesaurus encodes concepts from different vocabulary sources with Concept Unique Identifiers (CUIs). CUIs map concepts and terms across multiple terminologies. Likewise, the UMLS Semantic Network® compiles semantic relations (e.g. IS_A or CAUSED_BY) from the source ontologies. Terms in the UMLS are classified in 134 semantic types (STYs): e.g. *diabetes* is a Disease or Syndrome, and *fever* is a Sign or Symptom. UMLS semantic types are clustered into 15 semantic groups. For example, the group DISO contains various types of health conditions such as diseases, injuries, symptoms or findings. The UMLS MetaThesaurus contains 349,760 distinct French terms in the 2017AA version (4,011 of the type Sign or Symptom). Terms are often nested (e.g. *heart failure*) and variation includes derivation (*heart* vs. *cardiac*), compounding (*cardiovascular*), abbreviation (*MI*, for *myocardial infarction*) and lay terms (*heart attack*).

The above description shows that the number of different concepts and terms in this domain is larger than in usual dialogue systems. In such a context, we needed

[7] We thank the doctors from the Angers Medical School who recorded the sessions. Since the partner team collected the data, we do not know the number of recorded actors.

[8] `http://anglaismedical.u-bourgogne.fr/`

to provide the system with resources for concepts and words so that it knows about the domain and can handle term variation to interact adequately in it.

We developed for this purpose a *linguistic* and *termino-ontological model*, which hosts structured thesauri with linguistic, terminological and ontological knowledge. It is schematised in the central block of Figure 6 (left and right panels, respectively).

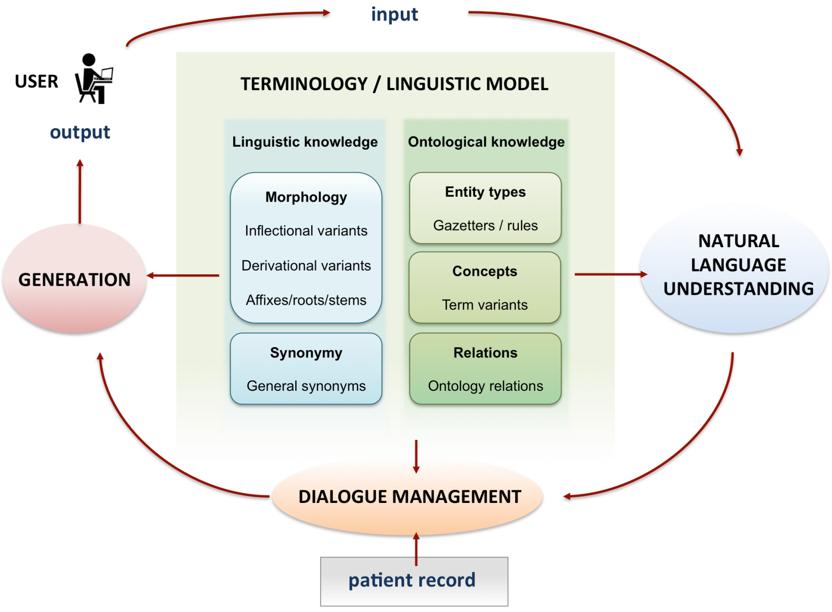


Fig. 6: Linguistic and termino-ontological model at the core of system architecture

With regard to the different language versions of the system, processing term variants is more challenging for French or Spanish, due to the higher number of verb forms or gender variants compared to English. For example, we needed resources for gender and number agreement to generate grammatically correct replies; accordingly, these resources are larger for the versions in French and Spanish.

3.3.1 Overall description

The variability of natural language expressions calls for *linguistic knowledge*. This includes word-level information: morphological information such as inflection (e.g. *kidney* ↔ *kidneys*), derivational variants (e.g. *surgery* ↔ *surgical*), affixes and root elements (e.g. *disease* ↔ *-pathy*), and synonyms (e.g. *operation* ↔ *surgery*).

Termino-ontological knowledge defines the relations and concepts that are useful for the system to interact in the domain. The structure of thesauri is similar to that in the UMLS Metathesaurus, and is organised around entity types, terms linked to concepts, and relations between these concepts.

We distinguish *entity types* (semantic classes of the domain defined for the task,

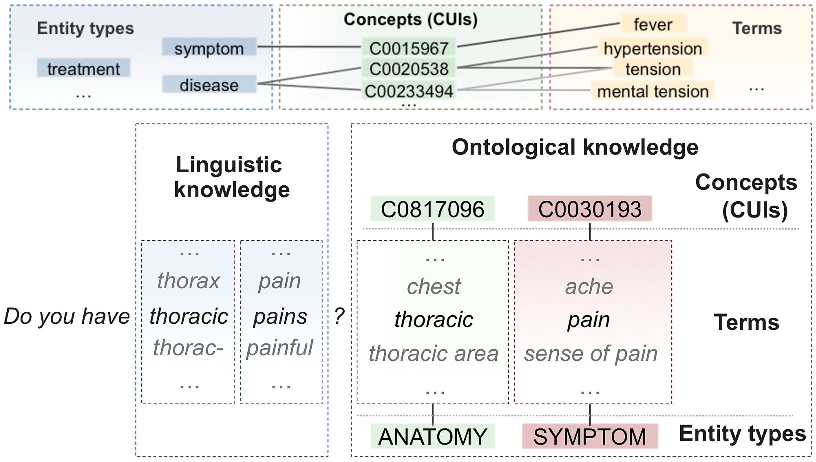


Fig. 7: Relation between entity types, concepts and terms. Linguistic knowledge processes affixes or morphological variants; ontological knowledge classifies terms into entity types; UMLS terms are indexed by Concept Unique Identifiers (CUIs)

e.g. label `treatment`) and *concepts* (conceptual items related to entity types). *Terms* refer to *concepts*, and concepts are classified with *entity type* labels (Figure 7). For each concept, termino-ontological knowledge provides one or more terms, which are used to handle term variation (e.g. *hypertension* ↔ *high blood pressure*).

The *linguistic knowledge* is instantiated through language resources such as dictionaries; it has no direct link to concepts or entity types. The *termino-ontological knowledge* is typically instantiated by domain terminologies. Accordingly, we use different linguistic and terminological resources. There is a separate lexicon file for each component (e.g. a file for synonym variants and another for derivational variants in the linguistic model). Table 2 shows the types of variation phenomena and the resources needed for the generation, entity linking and normalisation steps. For instance, the generation step (first pane of the table) uses information stored in both the linguistic and termino-ontological models (this is detailed in §4.5). Here, linguistic knowledge consists of morphological information: gender, number, and part-of-speech, as well as correspondences between specific verb forms (e.g., *has* ↔ *have*). Termino-ontological knowledge maps scientific and lay term variants (e.g., *per os* ↔ *oral*). Likewise, the entity linking and normalisation step (second pane of Table 2) uses linguistic knowledge to manage linguistic variation (see §3.3.3) and termino-ontological knowledge to manage terminological variation (see §3.3.4).

To build these lexicons, we extracted semi-automatically terms and ontology relations from the UMLS where possible. Due to the large size of the UMLS, we used the subset of its terminologies and semantic types that were relevant for our task; for example, we did not need entity types such as Regulation or Law (STY T089). We also used the National Agency for the Safety of Medicines list,[9] and extended

[9] `https://ansm.sante.fr/`

Generation	**Morphology** (LINGUISTIC KNOWLEDGE)	
	1.1. Gender/number/PoS data	*antihypertensive*, .N:ms (noun, masculine, singular)
	1.2. Verb correspondences	*[the patient] has* ↔ *[I] have*
	Concepts (ONTOLOGICAL KNOWLEDGE)	
	1.3. Scientific/lay correspondences (with CUIs)	*high blood pressure* ↔ *hypertension* (C0020538)
	1.4. Scientific/lay correspondences (no CUIs)	*per os* ↔ *oral*
Entity linking and normalisation	**Morphology** (LINGUISTIC KNOWLEDGE)	
	2.1. Inflectional variants	*[the patient] coughs* ↔ *[you] cough*
	2.2. Derivational variants	*impare* ↔ *imparement*; *fever* ↔ *feverish*
	2.3. Roots/affixes/stems	*append-* ↔ *appendix*
	Synonymy	
	2.4. General synonyms	*walk* ↔ *stroll*
	Concepts (ONTOLOGICAL KNOWLEDGE)	
	2.5. Terms with UMLS CUIs:	
	2.5.1. Anatomy	*thorax* ↔ *chest* (C0817096)
	2.5.2. Diseases and symptoms	*kidney disease* ↔ *nephropathy* (C0022658)
	2.5.3. Pharmacologic substances	*pain killers* ↔ *analgesics* (C0002771)
	2.5.4. Surgical and therapeutic proc.	*appendectomy* ↔ *appendix excision* (C0003611)
	2.6. Terms without UMLS CUIs:	
	2.6.1. Symptoms (verbs/idioms)	*to perspire* ↔ *perspiration, sweating*
	2.6.2. Other terms	*tension problems* ↔ *high blood pressure*
	Ontology relations	
	3. Relations between UMLS CUIs:	
	3.1 CHILD_OF	*hypertension* < *cardiovascular disease*
	3.2. Procedure–Disease	*inguinal herniorrhaphy* HAS_PROCEDURE_MORPHOLOGY *hernia*
	3.3. Procedure–Anatomy	*appendectomy* HAS_PROCEDURE_SITE *appendix*
	3.4. Disease–Physiological function	*shortness of breath* ↔ *breathing*

Table 2: Types of resources used in each processing step for managing linguistic and terminological variation (and examples of each type), listed according to linguistic aspects (morphology and synonymy) and termino-ontological aspects (concepts and ontology relations)

the lexicons as needed for our task. We built two types of lexicons: lexicons with terms recorded in the UMLS (with a CUI), and other lexicons that we created with terms not recorded in the UMLS. We explained the methods to collect them in a previous work (Campillos-Llanos *et al.* 2016), and we briefly describe them below.

3.3.2 Entity types

Vocabulary lists to label domain and miscellaneous entities (both mono- and multi-word items) amount to 161,878 items. Table 3 shows the correspondence between patient record sections, entity types, NLU labels, the source of the data, and the sizes of the associated term lists. The counts presented do not aggregate typographical variants; e.g. *anti-hypertensive* and *antihypertensive* are counted separately. We provide the source used for extracting some domain terms from the UMLS (column *Source*): codes are UMLS semantic types (STYs; e.g. T184 for Sign or Symptom) or UMLS semantic groups. Our list of diseases has items of all DISO STYs except Sign or Symptom (T184). We specify in italics the Medical Subject Headings (MeSH) code where this terminology was used: e.g. *D27.505* (Pharmacological action) for drug therapeutic classes. When there was not an UMLS semantic type for the entity type needed, we extracted terms by using regular expressions. For example, to collect the list of allergies, we applied the expression *(allerg|intolér|réaction)* on UMLS terms of type Disease or Syndrome (T047), and then manually revised results. Some lists come from the same STY, but semiautomatic and manual methods were used. For example, we extracted terms for gynecological and obstetric events from the type Disease or Syndrome. During development, after users interacted with the system, we observed that some question types were related to obstetrics entity types (e.g. *miscarriage*), and others, related to gynecologycal events (e.g. *abundant menstruation*). Hence we needed to distinguish between both of them and refined the scheme of entity types.

3.3.3 Managing linguistic variation

We address two types of linguistic variation: morphological variation (inflection and derivation) and synonymy. The former involves dealing with inflectional variants (e.g. *to sweat* and *sweating*) (Table 2, 2.1). We manage this through a general-language inflectional dictionary—we used DELA-type electronic dictionaries for French (Courtois 1990). Also, derivational variants may occur (e.g. *impairment* and *impair*; Table 2, 2.2). We deal with them through deverbal nouns collected from a specialised general lexicon—for French, VerbAction (Hathout *et al.* 2002). Derivational variants of medical terms (e.g. *fever* ↔ *feverish*, *thorax* ↔ *thoracic*) come from the UMLF lexicon (Zweigenbaum *et al.* 2005). Synonym relations (e.g. *walk* and *stroll*) are processed by means of general synonym lexicons (Table 2, 2.4). For French, we reused the dictionary applied in a previous project (Rosset *et al.* 2008). This open-domain dictionary contains entries whose values are synonym words: e.g. `opération|manipulation;action;intervention`.

Record section	Entity type	Internal label(s)	Example	Source	# terms
Medical history and/or symptoms	Allergies	`allergy`	*photoallergy*	T047	125
	Anatomy (n / adj)	`anatomy / anatomy_adj`	*thorax, thoracic*	ANAT	17,752
	Anesthesias	`anesthesia`	*epidural*	T061	388
	Bacteria	`bactery`	*coccus*	T007	4,118
	Body substance	`body_subs`	*blood*	T031	81
	Disease (spec.)	`disease_spec`	*hypertension*	DISO (excl. T184)	87,958
	Disease (gen.)	`disease_gen`	*disease*	-	18
	Gynecological event	`antec_gyn`	*menstruation*	T047	153
	Obstetric event	`obstetr`	*cesarean*	T047, T061	704
	Physiological function (n / vb)	`physiol / physiol_vb`	*digestion, to digest*	T039	139
	Surgical procedure (spec.)	`surgery_spec`	*graft*	T061, *E04*	2,649
	Surgical procedure (gen.)	`surgery_gen`	*surgery*	-	33
	Symptoms (n / vb)	`symptom / symp_vb`	*bleeding, to bleed*	T184	5,874
	Virus	`virus`	*parvovirus*	T005	2,564
	Descriptions of conditions:				
	Changes	`changes`	*aggravated*	-	269
	Colours	`colour`	*yellow*	-	61
	External characteristics	`desc_ext`	*bloody*	-	78
	Intensity/severity	`intensity`	*violent*	-	73
	Irradiation of pain	`irradiation`	*to irradiate*	-	27
	Onset type	`onsettype`	*progressive*	-	43
	Other features	`descript`	*stabbing*	-	165
	Volume	`volume`	*thick*	-	21
Miscellanea	Expressions of duration	`duration`	*for 2 hs*	-	50
	Expressions of frequency	`frequency`	*never*	-	135
	Expressions of manner	`manner`	*abnormally*	-	31
	Expressions of quantity	`quantity`	*many*	-	15
	Relative spatial position	`position`	*right*	-	34

Table 3: Correspondence between VP record sections, NLU labels and lists, and sources

Lifestyle	Addictive behaviour	`drug_addic`	*drug addict*	-	15
	Addictive substance	`drug`	*marijuana*	-	68
	Alcoholic behaviour	`alcoholism`	*drinker*	-	27
	Alcoholic beverage	`alcohol`	*wine*	T168	35
	Daily activities and acts	`act`	*to wash*	T056	118
	Diets	`regime`	*diet*	T061	38
	Food	`food`	*meat*	T168	1,125
	Recreational activities (n / vb)	`recr_act` / `recr_act_vb`	*swimming* *to swim*	T056	229
	Smoking behaviour	`smok_addic`	*smoker*	-	13
	Smoking products	`smok_prod`	*cigarette*	-	7
Treatments	Efficiency of treatment	`efficiency`	*efficient*	-	48
	Galenic form	`galen_form`	*pill*	-	98
	Medical drugs	`medicament`	*aspirin*	T121, *D27.505*	22,750
	Mode of administration	`mode_administ`	*orally*	T061	66
	Treatments	`treatment`	*lavage*	T061	1,270
	Therapeutic class	`therap_class`	*painkiller*	T121, *D27.505*	2,345
Exams and analyses	Diagnostic procedures	`analysis`	*radiography*	T060	2,618
	Exams or surgeries	`anal_surg`	*colonoscopy*	T060,T061	41
	Clinical/lab. tests	`lab_proc`	*blood count*	T059	4,439
Other record	Animals/pets	`animal`	*dog*	-	50
	Findings	`circums`	*physical effort*	T033	429
	Emergency admission	`urgence`	*emergency*	-	8
	Hospitalisation event	`hospitaliz`	*hospitalised*	-	8
	Medical devices	`medic_dev`	*pacemaker*	T074	1,775
	Medical spec. / doctors	`doctor`	*urologist*	*H02*	142
	Transfusions	`transfusion`	*autotransfusion*	T061	38
	Trips abroad	`travel`	*travelling*	-	25
	Vaccines	`vaccine`	*vaccination*	T121	91
TOTAL					161,878

Table 3: Correspondence between VP record sections, NLU labels and lists, and sources (*cont.*)

3.3.4 Managing terminological variation

Terms referring to the same concept For terms recorded in the UMLS (Table 2, 2.5), we can map term variants associated to the same concept by using their CUI: e.g. *high blood pressure* and *hypertension* (CUI C0020538). Term variants come from the following UMLS types and groups (McCray, Burgun and Bodenreider 2001):

- Anatomic entities (Table 2, 2.5.1): ANAT semantic group
- Health states (i.e. diseases/symptoms, Table 2, 2.5.2): DISO semantic group
- Pharmacologic substances (Table 2, 2.5.3): T121 semantic type
- Therapeutic/diagnostic procedures (Table 2, 2.5.4): PROC semantic group

Terms not recorded in the UMLS need special processing (Table 2, 2.6). For example, the UMLS lacks French verbs for symptoms (e.g. *to perspire*), which can not be mapped to noun forms (e.g. *sweating*, C0038990). We created ad-hoc lists to link them, gathering single- and multi-word verbs/idioms and lemmatised forms. We collected them manually and iteratively: after users interacted with the system, we analysed interaction logs and added items to lexicons. With a similar procedure, we collected a secondary list for remaining lay variants (e.g. *per os* ↔ *oral*).

When other methods fail, medical affixes and roots/stems (Table 2, 2.3) help match terms with no UMLS relation. To build these resources, we adapted the lexicon of affixes and roots of DériF, a morphosemantic linguistic-based parser for processing medical terminology in French (Namer and Zweigenbaum 2004). This analyser decomposes terms into morphological constituents and classifies them into domain semantic types. We also translated to French some neoclassical compounds (i.e. Latin or Greek prefixes, roots or suffixes, e.g. *amygd-*, ‘tonsil’) from the UMLS Specialist lexicon® (McCray, Srinivasan and Browne 1994). This lexicon is an English dictionary (over 200,000 entries) gathering biomedical terms and frequent words. Each entry records syntactic, morphological and orthographic information.

Using hierarchical relationships Some dialogue contexts involve concepts with a different degree of specialisation to that in the record. In these cases, we use UMLS relationships to map general to specific concepts (or vice versa), especially relations from SNOMED CT (Table 2, 3). The Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) (Donnelly 2005) is a clinically-oriented multilingual terminology distributed by the International Health Terminology Standards Development Organisation (IHTSDO). SNOMED CT gathers codes, hierarchical concepts, relations between them, and descriptions, and it is included in the UMLS.

Namely, UMLS CHILD_OF (CHD) relations are used to map a type of disorder (e.g. *cardiovascular disease*) to a specific disease in the record (e.g. *high blood pressure*). Relationships are also used when term variants fail. Terms referring to classes of disorders with the pattern `disease_gen` + `anatomy` are related to their anatomical site by SNOMED CT relation HAS_FINDING_SITE: e.g. *kidney disease* and *glomerulosclerosis*. Other SNOMED CT relations are used: HAS_PROCEDURE_SITE, to map anatomic terms and surgical procedures with pattern `surgery_gen` + `anatomy` (e.g. *appendectomy* ↔ *appendix*); and HAS_PROCEDURE_MORPHOLOGY or

Step	Resource	Variants	Min	Max	Mean	Entries/CUIs
Generation	1.1. Gender/number/PoS data	1,327,469	1	1	1.00	747,234
	1.2. Verb correspondences	48,432	1	1	1.00	24,830
	1.3. Scient./lay corr. (with CUIs)	23	1	9	5.20	4
	1.4. Scient./lay corr. (without CUIs)	58	1	1	1.00	38
Entity linking and normalisation	2.1. Inflection	631,035	1	61	7.96	91,569
	2.2. Synonyms	18,657	1	143	13.50	15,048
	2.3. Derivational variants	20,045	1	9	2.56	8,008
	2.4. Terms with CUIs:					
	2.4.1. Anatomy	18,143	1	29	3.15	3,749
	2.4.2. Diseases/Symptoms	367,887	1	34	2.86	58,853
	2.4.3. Pharmacologic substances	22,283	1	23	3.10	6,465
	2.4.4. Surg./therap. procedures	130,493	1	24	2.63	20,975
	2.5. Terms without CUIs:					
	Symptoms (vbs./idioms)	815	1	36	14.36	58
	Other terms	183	1	22	7.56	25
	Roots/affixes/Stems	712	1	12	2.23	319
	3. Relations between CUIs:			# Pairs of concepts (CUIs)		
	CHILD_OF			170,571		
	Procedure - Disease			11,854		
	Procedure - Anatomy			95,744		
	Disease - Phys. function			8,144		

Table 4: Resources for managing linguistic and terminological variation

HAS_DIRECT_MORPHOLOGY, to map entities referring to surgeries with the structure `disease_spec` + `surgery_spec` (*hernia* ↔ *inguinal herniorrhaphy*).

To relate symptoms or disorders to physiological functions (e.g. *dyspnea* and *to breathe*), we extracted lists of correspondences between these types of entities from UMLS terminologies—namely, ICD10, MeSH and SNOMED—together with hierarchical relationships between concepts of these types.

Table 4 reports the size of the resources for managing terminological and lexical variation: number of variants, minimum, maximum and mean values per CUI or lexical entry (mono- and multi-word items), and number of lexical entries or CUIs in each resource (for relations between CUIs, we give the number of related pairs).

4 Implementation

In this section, we first describe the general architecture of the system (§ 4.1). We then detail the resources and processes used in the NLU stage (§4.2 and §4.3), the dialogue manager and patient record querying (§4.4) and the generation step (§4.5).

4.1 Architecture of the dialogue system

The architecture of the system (Campillos-Llanos *et al.* 2015) is based on the modular schema of the RITEL interactive question-answering dialogue system for open-domain information retrieval (Rosset *et al.* 2006). The RITEL platform is an infrastructure for developing dialogue systems. We adapted the NLU engine and the processing functions for dialogue management and generation. Our contributions are the lexicon and models for the VP dialogue task. Our system has these modules:

- The *terminology and linguistic management module* provides the termino-ontological and linguistic knowledge needed by the various components.
- The *natural language understanding (NLU) module* analyses the user input. It recognises medical entities and question types, and includes spelling correction to deal with errors in user input. The user input is a turn and consists of one or more utterances. For instance, *Hello. How are you doing?* is a turn and consists of two utterances.
- The *dialogue manager* interprets the results of the NLU. Based on the entity and question labels produced by the NLU, rules are applied to determine the semantic frame that best represents the type of user question and drives the construction of a query to the record. The required information is looked for in specific sections of the record. The dialogue manager then passes the query results and a suitable reply type to the generation module.

  The dialogue manager takes into account and updates a dialogue history, using a shallow implementation of the information state approach (Traum and Larsson 2003). This history keeps track of the interaction at each move, and is used to process ellipsis and anaphora (see examples in Figure 1). This is an important device to manage longer, more natural conversations than what a series of independent question-answer turns would provide.
- The *natural language generation module* creates the output utterances. Instead of using predefined replies, the system relies on templates to generate new answers according to the contents of the patient record.

Figure 8 illustrates how an input question is analysed and processed to output a reply according to the VP record. After the spelling correction, the NLU module detects the terms in the user input and annotates the question and entity types. Then, the dialogue manager processes the semantic frame filled with the entity types in the input. The terminology management module looks for the corresponding data in the VP record, performing entity linking or normalisation if needed. If these data are found, the lexical module looks for a lay variant of the term (e.g. *appendicitis operation* for *appendectomy*). Lastly, the generation module applies the type of reply that corresponds, and the record items are instantiated in the generation template.

4.2 NLU: Spelling correction

For each out-of-vocabulary word, we attempt spelling correction. To handle common types of misspellings, we implemented a correction algorithm based upon Norvig (2007). We use morphological information, word length and corpus frequency of each word in the system together with edit-distance metrics to choose the most likely correction. The spelling corrector relies on a dictionary gathering all mono-word terms in the system. Each word form corresponding to the same

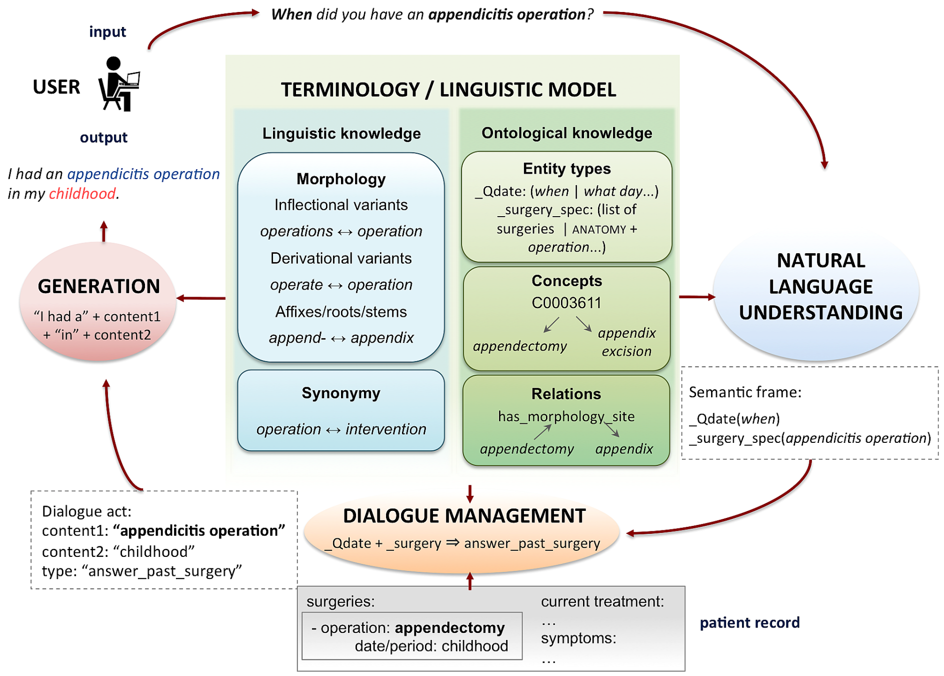


Fig. 8: Example of question-reply processing

lemma has its own entry (e.g. *tension* and *tensions*). The current version contains 659,720 word forms.

We used a domain corpus to develop the dictionary of the spelling corrector. The VIDAL® company provided for the project its dictionary of medical drugs, which is a reference book used by health professionals and is copyrighted (therefore, not freely available).[10] Texts include data on common disorders and medications in French, and the corpus size amounts to 7,678,363 tokens.

Spelling correction is a useful component but is not the focus of the present paper, hence we do not expand further on its description.

4.3 NLU: Entity recognition and semantic annotation

The NLU module performs the following tasks:

- Entity recognition and semantic annotation: we use rules and domain lists and gazetteers semi-automatically curated for each entity type. More detail on these lists were provided in §3.3 and Table 3.
- Dialogue act and question type annotation: we use rules to label conversa-

[10] If free texts are needed for a similar domain, a corpus of the European Medicines Agency is available at: `http://opus.nlpl.eu/EMEA.php`

```
# Grammar and rules for acknowledging
_acknowledging: ( (?: ^ok | ^okay | all right | yes ) (?! "?" ) );

# Grammar and rules for greeting
_greeting: ( how are you doing | how are things going | what's up )

# Contexts and words for time expressions (hours)
&hour: half-hour | half-hours | _~hour | _~minute | noon | midnight |
o' clock | pm | am | p.m. | a.m. ;

# Contexts and words for time expressions (parts of the day)
&part_of_day: (<! good) ( morning | night | evening ) | diurnal ;

# Grammar and rules for questions on hours
_Qhour: ( (?: what | which ) .{0,2} time | (?: at | around ) about?
(?! between | from ) %cardinal(0,24) &hour | &part_of_day );

# Specific terms of surgeries (appendectomy, tonsillectomy...)
&surg_spec_trms: include LIST_OF_SURGERIES ;

# General, unspecified terms referring to surgical procedures
&surg_gen_trms: intervention | operation | procedure | surgery ;

# Terms of anatomy (appendix, chest, knee, ligaments, tonsils...)
&anatomy: include: LIST_OF_ANATOMIC_ENTITIES;

# Grammar and rules for specific types of surgery procedures
_surgery_spec: ( &surg_spec_trms | &surg_gen_trms of &surg_spec_trms |
(?: &surg_gen_trms | _~operate | _~remove ) .{0,4} &anatomy ) ) ;

# Specific terms of disorders (cancer, diabetes, hypertension...)
&dis_spec_terms: include LIST_OF_DISEASES;

# General, unspecified terms of disorders
&dis_gen_terms: disease | disorder | illness | pathology ;

# Grammar and rules for specific types of disorders
_disease_spec: ( (?: &disease_gen_terms .{0,5} &anatomy ) |
&dis_spec_terms );
```

Table 5: NLU rules (actual code) for dialogue acts (greetings and acknowledging), questions on hours, and entity types (diseases and surgeries). Labels start with the _ character; & indicates a sub-expression that can be reused in different locations; | indicates alternation; ∧, start of string; <!, left negative lookahead; ?!, right negative lookahead; ?:, non-substituting grouping; .{0,n}: 0-n matching; _∼, word lemma

tional dialogue acts (e.g. greetings) and classify question types (e.g. *yes-no* questions bear label Qyesno).

In this step, we use Wmatch rules (van Schooten *et al.* 2007; Galibert 2009). Wmatch is a regular expression engine of words for natural language processing. It

			Question structure	
			***Wh-* type**	***Yes-No* type**
Ent. type	**General**		*What are your symptoms ?*	*Do you have any symptoms?*
	Specific	**Class**	*How do you breathe?*	*Can you breathe well?*
		Subclass	*What type of breathlessness?*	*Are you breathless?*

Table 6: Types of input questions

uses domain lists to detect words in user input and allows defining local contexts for matching and semantic categorisation. Each of the 149 entity types, dialogue acts and question types is defined by a grammar that is expressed by a combination of abstract rules and gazetteers. Each grammar generates a complex graph that is used ultimately at run time. Table 5 shows sample rules. Two computational linguists, with the expertise of a senior researcher, developed the rules in an iterative process of analysing interaction logs and refining matching contexts.

With regard to medical entity types, we distinguish two levels of specialisation: general (the top-level entity types of the domain, such as `surgery_gen`) and specific (the descendants of these top-level types, such as `surgery_spec`). Patrick and Li's taxonomy also differentiates general and specific clinical questions (Patrick and Li 2012), but we did not adapt it to our knowledge model for the dialogue task due to the different application setting.

We consider two variants of question types: *Wh-* questions (*open questions*) and *yes-no* questions (*polarity questions*) (Quirk *et al.* 1985). This distinction is needed to determine the type of reply (i.e., *yes-no* questions are replied with *yes* or *no*).

The resources used in the NLU step vary according to the type of entities in input questions (see Table 6):

- *Questions on general entities*: Lists of *top-level* entities (i.e. towards the top of the hierarchy or ontology) referring to a entity type in the domain: e.g. *operation* belongs to the entity type `surgery_gen`, and *disorder*, to the entity type `disease_gen`.
- *Questions on specific entities*: *specific* entities referring to a more detailed entity type in the domain. Two types of queried entities may appear depending on the type of question topic:
  - — *Classes of entities*: e.g. *cardiovascular disease* (`disease_spec`).
  - — *Subclasses of entities*: e.g. *appendectomy* (`surgery_spec`) or *hypertension* (`disease_spec`).

In the NLU step, lists and rules aim at balancing both precision and recall to consider different term and question structures or spelling variants. To increase precision, we needed comprehensive lists of terms and rules defining precise matching contexts. To improve recall, we expanded term lists (e.g. by including frequently misspelled words) and relaxed the context of some rules—the less specific the context, the higher the recall. During system development, we removed noisy terms in lists and fine-tuned greedy matching rules.

4.4 Dialogue manager and patient record querying

At each dialogue move, the dialogue manager interacts with the lexical modules. First, the dialogue manager processes user input according to the semantic frame from the NLU step. In addition, the information state module updates the input content representation dynamically according to the current dialogue state. The reference of an anaphoric pronoun or an elliptic element is interpreted according to the previous dialogue state. For example, in the sample dialogue of Figure 1, the system interprets the ellipsis of the medical term in *since when?* as the symptom expressed in the previous reply (*shortness of breath*). This allows the system to manage the semantic interpretation of user input in context.

To query the record, inflected forms of entities are transformed to a base or canonical form (e.g. the singular noun or the infinitive verb form) by using the lexicons in the linguistic model. Medical entity verbs (e.g. *Have you bled?*) undergo some steps of lemmatisation (*bled* → *bleed*) before the base form is mapped to any variant (e.g., *bleed* → *hemorrhage*). Multiword entities also need another step to remove some pronouns and obtain a canonical form: e.g. *Respirez-vous avec difficulté ?* ('Are you breathing with difficulty?') is reduced to the base form *respirer avec difficulté* ('breathe with difficulty'); then, this form can be mapped to a mono- or multi-word variant term in the patient record (e.g. *shortness of breath* or *dyspnoea*).

In the patient record query step, postprocessed entities are dynamically looked for in the record. The dialogue manager uses the entity type to restrict the search for data in the corresponding record section. For example, a question on a disease is looked up in the section concerning disease history.

There is a continuum between types of entities and questions types (as exposed in §4.3). Their nature (general or specific) affects the size of processes and resources for querying the record. At one extreme, questions on general entities only require an accurate identification of the entity type. For example, a question such as *What diseases do you have?* requires identifying *diseases* as a generic term (label `disease_gen`). At the other extreme, questions on specific entities also demand *entity linking* (also called *entity normalisation*) to check whether the input entity and that in the record refer to the same concept. For example, a question such as *Do you have tension problems?* requires labelling *tension* as `symptom`, and managing term variants when checking these data in the record (e.g. *tension problems* ↔ *high blood pressure*). Questions on a class of entities require matching this class with any of its subclasses in the record. If a user asks a question such as *Do you have cardiovascular diseases?*, which contains a term referring to a broad class, we need to map it to a specific disease in the record (e.g. *high blood pressure*, a subclass of *cardiovascular disease*). To do so, we use ontological relations.

Algorithm 1 Pseudocode of function to match terms through UMLS CUIs.
The function returns True when an input term and a term in the patient record refer to the same UMLS concept. Dictionaries in the termino-ontological model are selected according to a semantic code (ANAT, DISO, PROC or T121) corresponding to the input entity type.
The linguistic model model is used to get the lemma of the input word form.

```
function MATCH_TERM_THROUGH_CUI(input_term, record_string, semantic_code)
    # Lowercase input term                                   NORMALISATION
    input_term_lc = LOWER(input_term)
    # Select dictionary of terms                  USE TERMINO-ONTOLOGICAL MODEL
    # Default value of semantic code
    semantic_code = DISO
    if semantic_code = ANAT then
        term_dic = list_variants_anat_CUI
    else if semantic_code = DISO then
        term_dic = list_variants_diso_CUI
    else if semantic_code = PROC then
        term_dic = list_variants_proc_CUI
    else if semantic_code = T121 then
        term_dic = list_variants_T121_CUI
    end if
    # Check common CUI in list of variants
    if HAS_COMMON_CUI(input_term_lc, record_string, term_dic) then
        return true
    end if
    # Check the lemma of the input word                 USE LINGUISTIC MODEL
    lemmas_list = GET_LEMMA(input_term_lc)
    for i=1,#lemmas_list do                   USE TERMINO-ONTOLOGICAL MODEL
        if HAS_COMMON_CUI(lemmas_list[i], record_string, term_dic) then
            return true
        end if
    end for
    return false
end function
```

Methods for entity linking use exact or approximate match (Levenshtein 1966) or any of the resources defined in the termino-ontological model. The specific lexicons and/or ontology knowledge to be used rely on the entity type of each term. Terms whose entity types are related to pathologies (e.g. label `disease_spec` or `symptom`) are looked up in lexicons of term variants extracted from the UMLS DISO group. That way, *hypertension* can be mapped to *high blood pressure* or *hypertensive disorder*. Likewise, terms belonging to procedure entity types (e.g. *appendectomy*, label `surgery_spec`) are looked up in lexicons with variants extracted from the UMLS PROC group. The input terms to be matched with terms in our lexicons belong to the same entity type; variants are not expected to be found among terms of other entity types. This restriction of the search space speeds up the dictionary look-up process. By focusing on the relevant parts of the record, the cor-

Algorithm 2 Pseudocode of function to query surgery terms in the patient record. The function returns True when an input term is found in the patient record. Dictionaries and ontology relations are used from the termino-ontological model. Function MATCH_TERM_THROUGH_CUI (see algorithm 1) is used to match terms through UMLS CUIs.
The linguistic model is used to match terms through affixes and roots.

```
function CHECK_FOR_SURGERY(input_term, record_string)
    # Lowercase input term                                          NORMALISATION
    input_term_lc = LOWER(input_term)
    # Exact or approximate match
    if EXACT_OR_APPROX_MATCH(input_term, record_string) then
        return true
    # Dictionary (terms with CUIs)                      TERMINO-ONTOLOGICAL MODEL
    else if MATCH_TERM_THROUGH_CUI(input_term_lc, record_string, PROC) then
        return true
    # Dictionary of terms without CUIs
    else if MAP_TERM(input_term_lc, record_string) then
        return true
    # Ontology relations (procedures ↔ anatomy)
    else if MAP_PROC_ANAT(input_term_lc, record_string) then
        return true
    # Ontology relations (procedures ↔ disorders)
    else if MAP_PROC_DISO(input_term_lc, record_string) then
        return true
    # Match through affixes and roots                      USE LINGUISTIC MODEL
    else if MATCH_THROUGH_AFFIX(input_term_lc, record_string) then
        return true
    end if
    return false
end function
```

rection of answers is also expected to increase. Algorithms 1 and 2 are pseudocode examples of how these queries are implemented.

In this step, the correction of answers depends on the ability of the system to map input terms to items in the patient record. This in turn depends on the coverage and quality of the linguistic and termino-ontological resources of the system.

4.5 Generation

Resources for generating replies cover three types of information:

- Linguistic data for gender/number agreement: e.g. *fever* is feminine in French. We use DELAS-type (Courtois 1990) dictionaries with inflectional information (Table 2, 1.1).
- Correspondences between 3rd and 1st person verb forms, to output the content expressed in the record (in 3rd person) with the patient's viewpoint (1st

person): e.g. *The patient has a fever* → *I have a fever*. We clustered pairs of verb forms from the mentioned dictionaries (Table 2, 1.2).

- Lay variants of terms: e.g. *appendectomy* → *appendicitis operation*. These were selected by processing domain corpora of different degrees of technicality (Bouamor *et al.* 2016) and manual revision (Table 2, 1.3 and 1.4).

5 Evaluation methods and results

We present our evaluation goals and criteria (§5.1) and explain how we gathered evaluation data (§5.2). Next, we detail our evaluation methods and results for different aspects, and we end with a discussion of results (§5.8).

5.1 Overview of evaluation principles

One of the difficulties in evaluating dialogue systems lies in the lack of benchmarks and comparable or agreed standards (Paek 2001). Frameworks such as PARADISE (Walker *et al.* 1997) established a foundational methodology, especially with regard to distinguishing objective and subjective metrics—or *performance* and *usability* (Roy and Graham 2008). Human judgements on dialogue performance are thus relevant and necessary to complement other measures.

We designed and ran both quantitative and qualitative evaluations of system performance with a focus on its vocabulary coverage. Evaluating at these two levels provides us with an overall picture of how objective metrics reflect subjective assessments (Paek 2001). More specifically, we performed the following evaluations:

- A quantitative evaluation of the *natural language understanding* unit (§5.3).
- A quantitative evaluation of *dialogue management*, i.e. dialogue control and context inference (§5.4).
- A qualitative evaluation of the overall functioning of the system and of its usability (end-user satisfaction) (§5.5).
- A quantitative evaluation of the system's vocabulary coverage with regard to processing new cases (§5.6).
- A qualitative evaluation of vocabulary usage in the task (§5.7).

5.2 Collection of interaction data

During system development, we collected interaction data by having computer science students and researchers (n=32) interact with the system (3 VP cases) and evaluate it through an online interface and questionnaire.[11] For the evaluation presented here, in the following rounds of tests, medical students and doctors (n=39) interacted freely with the system and then evaluated it. We used 35 different VP cases; each case was tested by an average of 3.74 users (±2.8; minimum number of different users per case=1; maximum=13). We gave instructions concerning the

[11] http://www.audiosurf.net/pg_2018/select_case.php

types of dialogue acts the system can process (e.g. avoid instructions or out-of-task requests such as *Give me your telephone number*). Table 7 includes a sample of the instructions provided. Note that we did not provide examples of question formulations, removing the risk of priming effects.

You are a medical doctor and a 41 year old woman, married, arrives for a preoperative assessment.
You need to get information about the following aspects:

- Patient's demographic data: weight, height, profession, family...
- Lifestyle: eating habits and diets, addictions, physical activities...
- Medical history: allergies, diseases, surgery history, obstetric history, treatments...
- Patient's symptoms, treatments and observations

The system cannot reply (or replies wrongly) to:

- Instructions or prescriptions: e.g. *I will examine you*, *I will prescribe you a medicine*
- Complex or coordinated questions: e.g. *Do you smoke or drink?*, *How much and since when?*

Try asking your questions in a varied way and avoid abbreviations if possible.
You can fill in an evaluation questionnaire whenever you want by clicking on the button *Evaluate the system*.

Table 7: Instructions available on the evaluation interface

The data reported in this evaluation were collected between March 2016 and February 2018. During the development tests with a small set of cases, we collected from computer science students and researchers around 1,987 pairs of turns, i.e. user input and system reply (a total of 3,756 turns, 11,960 tokens in user input). Users with this profile have more knowledge of the human-computer interaction limits and helped us improving system's response behaviour. Dialogues they conducted provided us with variants of question types before real end-users evaluated the system. For the evaluation presented here, we gathered from 39 medical doctors a total of 8,078 turns in 131 interaction dialogues (21,986 tokens of user input and 21,921 tokens in replies). After manually inspecting our data, we removed 149 pairs of turns (3.7 per cent) corresponding to out-of-task questions (e.g. *What is your favourite colour?*) or declarative statements the system is not expected to answer (e.g. *Please give me reasonable replies*). We are aware that some types of declarative statements are important in doctor-patient interactions, e.g. to show empathy or counsel the patient (e.g. *You should stop smoking*). The current version of the system includes rules to process some declarative statements related to the patient's additive behaviour. Table 8 breaks down our evaluation data (all collected turns) corresponding to medical users. Note that in some cases the system did not reply due to processing errors of the dialogue manager.

Users	**#D**		**Turn pairs**			**Words**		
			#T	**#T/D**	**stdev**	**#W**	**#W/D**	**stdev**
39	131	Input	4,044	30.87	11.71	21,986	167.83	78.32
		Replies	4,034	30.79	11.70	21,921	167.34	78.46
		Total	8,078	61.66	11.70	43,907	335.17	78.39

Table 8: Data collected during the evaluation by medical doctors; *#D*: count of dialogues; *#T*: count of turns; *#T/D*: average turns per dialogue; *#W*: count of words; *#W/D*: average words per dialogue; *stdev*: standard deviation

5.3 Quantitative evaluation of NLU

5.3.1 Methods

The natural language understanding module analyses the user input and as a result provides labels for its semantic analysis. To evaluate it, we used the standard metrics of Precision, Recall, F-measure, and Slot Error Rate (SER) (Makhoul *et al.* 1999). These were computed by counting the number of correct and incorrect entity types labelled; for wrong labels, we counted insertions (I), deletions (D) and substitutions (S) (Table 9). We also computed the Sentence Error Rate (SeER), which is the proportion of sentences with at least one error. The definitions of all these metrics are given in Table 10. A better performance is normally reflected in higher P, R and F-measures and lower SER and SeER values.

We also evaluated the spelling correction module by counting the number of errors in dialogues. However, we found that most spelling errors affected grammatical words or items not requiring semantic annotation. We thus manually corrected the orthographic errors in the input and evaluated the NLU again. Our results showed that the F-measure did not largely improve (+0.4 per cent). This confirmed that spelling errors had a minor impact on the semantic annotation. Given those results and the space limits of this article, we do not report the evaluation of this module.

5.3.2 Results

We reported a preliminary evaluation based on 242 turn pairs in a previous article (Campillos-Llanos *et al.* 2016). Herein, we present the results based upon the data obtained to date from 39 medical users (around 4,044 turn pairs). Table 11 shows that the NLU achieves a high F-measure (95.8 per cent), balancing precision

Class	**Definition**		
Hypothesis (*Hyp*)	Total of correct and wrong labels in the system annotation		
Reference (*Ref*)	Total of correct and wrong labels in reference annotation		
Correct (C)	Number of correct labels in the hypothesis (*true positives*)		
	User input	Hyp	Ref
	How much fever?	`Qtemperature`	`Qtemperature`
Deletion (D)	Number of missing labels in the hypothesis (*false negatives*)		
	User input	Hyp	Ref
	Are you addicted to something?	Ø	`Qaddictions`
			(1 deletion)
Insertion (I)	Number of inserted labels in the hypothesis (*false positives*)		
	User input	Hyp	Ref
	What profession do you practice?	`Qactivity,`	`Qactivity`
		`act_recr_vb`	(1 insertion)
Substitution (S)	Number of replaced labels in the hypothesis		
	User input	Hyp	Ref
	Do you have psychological troubles?	`Qyesnosymptome`	`disease_spec`
			(1 substitution)

Table 9: Evaluation schema for the NLU module (with examples)

(96.8) and recall (94.9). Recall is lower than precision due to missing entity types and unannotated terms. The 6.1 per cent of Slot Error Rate implies that, on average, one entity type was incorrectly labelled every 16.4 labels. The 10.7 per cent of Sentence Error Rate means that, overall, one turn with incorrect labels occurred every 9.3 turns. Both these error rates are low. In terms of impact on dialogue flow, errors at this level made users reformulate their questions or change the topic of the dialogue.

Precision (P) is the ratio between the correct labels for entity types and all labels in the hypothesis:

$$P = \frac{C}{Hyp}$$

Recall (R) is the ratio between the correct labels in the hypothesis and the reference labels:

$$R = \frac{C}{Ref}$$

F-measure (F) is the harmonic mean of P and R; it is usually weighted with $\beta = 1$ and provides a global summary of the system performance:

$$F = \frac{(1+\beta^2)PR}{\beta^2 P + R} = \frac{2PR}{P+R}$$

The Slot Error Rate (SER) is similar to the Word Error Rate and is used in information extraction tasks (Makhoul *et al.* 1999). The SER is the total number of annotation errors (I, S and D) divided by the total count of labels in the reference (insertions are not considered in the reference, i.e. the denominator):

$$SER = \frac{S + D + I}{Ref}$$

The Sentence Error Rate (SeER) is the ratio between the sentences with at least one error and all of the correct sentences:

$$SeER = \frac{\#Wrong\ sentences}{\#Correct\ sentences}$$

The proportion (in per cent) of correctly-corrected spelling errors (CC) is calculated by dividing the number of CC errors and the sum of errors that were not corrected (NC), wrongly well-corrected errors (WC) and correctly-corrected errors:

$$Proportion\ of\ spelling\ errors\ (per\ cent) = \frac{CC}{(NC + WC + CC)}$$

Table 10: Evaluation metrics for the natural language understanding evaluation

5.4 Quantitative evaluation of dialogue management

5.4.1 Methods

To evaluate the dialogue manager, we adopted the framework explained by Dickerson and colleagues, who evaluated a system developed for the same task and domain (Dickerson *et al.* 2005). We analysed the logs of user-system interactions and manually classified each turn pair as *correct*, *incorrect*, *not understood* or *clarifica-*

Entities in reference	**Entities in hypothesis**	**Average ± stdev**	**Minimum**	**Maximum**
9,488	9,301	2.46 ±1.26	0	12
Correct	**Insertion**	**Deletion**	**Substitution**	**Total errors**
9,000 (94.9)	126 (1.3)	313 (3.3)	175 (1.8)	614 (6.5)
Precision	**Recall**	**F-measure**	**SER**	**SeER**
97.5	94.9	95.8	6.1	10.7

Table 11: Evaluation of the entity detection (NLU). We report the average number of entities per turn annotated in system hypothesis, standard deviation (*stdev*), minimum and maximum number of entities. Values for precision, recall, F-measure, SER and SeER are given in percentages, as well as figures between brackets

tion requests (Purver, Ginzburg and Healey 2003). We define a correct reply as that providing both: 1) a coherent answer with regard to the user question; and 2) correct information from the VP record. Conversely, incorrect replies are those not succeeding at providing a coherent answer, or those giving erroneous information. Three of these criteria also map to those applied by Traum and his team (Traum, Robinson and Stefan 2004): (1) Correct utterances correspond to Traum and colleagues' *Get response* and *Appropriate continuations*; (2) Clarification utterances correspond to *Request for repair*; (3) Incorrect utterances correspond to *Inappropriate response or continuation.* A computational linguist analysed interaction logs and classified them; then, a subset of 350 (8.6 per cent) turn pairs, which were hard to interpret and classify, was double-checked by a senior computational linguist; finally, a consensus was reached. We computed the percentage of turn-reply pairs for which the annotations made by one researcher were confirmed by the senior researcher at the consensus stage. The agreement between both linguists was of 93.6 per cent.

5.4.2 Results

An analysis of the user-system interaction logs was performed on the data collected from non-medical users in the development stage, and showed that 72.9 per cent of the replies were correct. The analysis of the interaction logs collected from medical users is shown in Figure 9 and Table 12. Segment labeled 1 in Figure 9 stands for correct replies, which represents 74.3 per cent; segment 2, the ratio of incorrect replies (14.9 per cent); segment 3, the proportion of not-understood replies (7.8 per cent); and segment 4, the ratio of clarification requests (2.9 per cent). Performance across

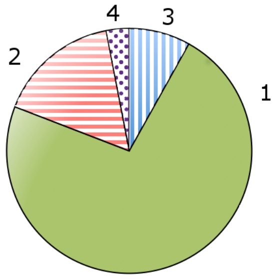


Fig. 9: Dialogue manager evaluation

Table 12: Dialogue manager evaluation (medical doctors)

	Count	Average per cent (stdev)	Min	Max
1. Correct	2,835	74.3 (±9.5)	53.6	93.8
2. Incorrect	628	14.9 (±6.3)	0.0	31.6
3. Not understood	315	7.8 (±5.3)	0.0	25.0
4. Clarification request	117	2.9 (±2.7)	0.0	11.5
Total evaluated	3,895	100		

VP cases varied (standard deviation, stdev, of 9.5) due to the different number of dialogues conducted with each case, and also in relation to the medical specialities of the cases. We obtained the best results (93.8 per cent) with a VP case suffering from diarrhea, and poor results (53.6 per cent of correct replies) with a postpartum case from the obstetrics speciality; however, both of these were tested by only one evaluator. In our error analysis of the logs of the postpartum case, we noticed that some of the evaluator's questions referred to the patient's newborn. The dialogue manager provided wrong replies because these question types did not refer to the patient's medical condition, but to that of her newborn, and the system could not distinguish them. Among incorrect replies, about 37.8 per cent were due to errors in the dialogue manager and 26.2 per cent were caused by unforeseen question types (e.g., we did not prepare rules for questions on the patient's blood group). Among the not-understood replies, 48.2 per cent were caused by unforeseen question types and about 10.2 per cent were caused by missing variants of questions.

5.5 *Qualitative evaluation of system performance and usability*

5.5.1 *Methods*

Right after users interacted with the system, they filled in a questionnaire with questions using a 5-point Likert scale. The survey addressed the following aspects:

- Global functioning: an overall assessment of system performance.
- Coherence: adequateness of system answers in relation to user input.
- Informativeness: satisfaction with the information provided by the system.
- User-understanding: degree of comprehension of system replies by the user.
- System-understanding: system's degree of comprehension of user input.
- Speed: system quickness in replying.
- Tediousness: verbosity of information answered by the system.
- Answer concision: quality of replies in terms of length.
- Naturalness of replies: realism of the utterances produced by the system.

5.5.2 *Results*

The 131 questionnaires collected from medical users scored highly the degree to which users understood system replies (64.1 per cent of evaluators assessed it *very*

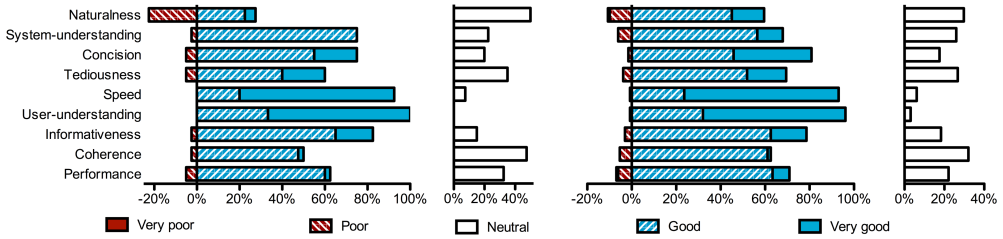


Fig. 10: Qualitative evaluation by non-medical users (left) and medical users (right)

good) and the speed in providing an answer (*very good*, 69.5 per cent). The following aspects were in general considered *good*: overall performance (63.4 per cent of users), informativeness (62.6 per cent), coherence of replies (61.1 per cent), system understanding of input (56.5 per cent), concision of replies (45.8 per cent) and their (absence of) verbosity or tediousness (51.9 per cent). The naturalness of replies was scored as *good* by 45.0 per cent of participants; 29.8 per cent gave a *neutral* score, and 9.9 per cent assessed it as *poor*. There is still room for improvement for this and other aspects; lower scores, however, represented only a small proportion of users.

Figure 10 depicts the results of the evaluation through the 40 questionnaires collected from participants with computer science backgrounds (left) and the 131 questionnaires filled by medical doctors (right). Assessments were rather similar excepting slight variations regarding informativeness and system understanding (slightly higher for computer science users) or naturalness (slightly higher for medical users). These differences might be due to the more strict criteria applied by medical users and to the improvements made to the system between evaluation rounds.

Users provided free comments concerning aspects to be improved. Table 13 shows some of them (translated from French). Several users commented upon difficulties in getting more details after a general question. Sometimes the record lacks detailed information: this raises the question of what the system should answer if the user asks for such missing information. For example, some users asked for the patient's disease or symptoms, and after the system replied, they wanted to know specific observations, which were not present in the record. In that situation, currently, the dialogue manager gives an explicit answer (*I cannot answer that question. This piece of information is not present in the record*). This does not always satisfy users due to the missing data or to the lack of naturalness of the reply (see Table 13). Because medical users need accurate information from the patient, we chose to give a neutral reply when no data are available. Likewise, some context processing errors have hindered the correct interpretation of questions. The first case requires to improve the ergonomy of the system, the latter case requires to improve its robustness.

Negative comments	Positive comments
Replies are very stereotyped (...) As soon as one goes out of the strict context of expected questions, the system is lost	*I have noticed its limitations, but it is often possible to reformulate to get a coherent answer since the system replies that it did not understand*
The patient always said "I cannot answer to that question [There is no information in the record]", which makes the dialogue less natural	*For some questions, the patient only replied almost always the same thing, but apart from that, the dialogue is natural and fluid, the patient understood many things.*
Sometimes not too much memory of previous question	*System replies are fine and make it possible a fluent interaction*
I didn't have the impression that it was possible to link several questions, that is, to clarify certain answers	*Very very coherent replies, some sentences where syntax was not completely correct (sometimes a verb is missing). The patient gives a lot of information anyway and the dialogue is fluid.*

Table 13: A selection of positive and negative user comments in the qualitative evaluation (translated from French).

5.6 Quantitative evaluation of vocabulary coverage

5.6.1 Methods

We assessed how robust our lexicons are by comparing them to domain data not used for developing the system. Because no preexisting library of VP cases existed in French, we used the most similar source we could find. We collected 169 cases from Epreuves Classantes Nationales ('National Classifying Tests', hereafter ECN), which are used to prepare exams in medical universities.[12] We used the description of the case, not the feedback for students. Table 14 shows a sample.

The procedure was as follows. We lowercased and tokenised the ECN texts; we removed numbers, dates, punctuation and stop words; and we expanded common abbreviations (e.g. *mg* → *milligrams*). Then, we compared the word types (i.e. different word forms, not tokens, which represent the occurrence of each type) in these texts against all the terms in the lexical resources used by the NLU, entity

[12] Freely available online at: `http://umvf.cerimes.fr/portail/ecn.php`

Monsieur B., 71 ans, consulte pour des douleurs abdominales et des épisodes de constipation. Il présente également des épisodes de selles molles. Son médecin traitant lui a parlé d'intestin irritable, ce qui ne l'a pas vraiment rassuré. Il est comptable retraité. Il a des antécédents d'hypertension artérielle et d'artérite pour lesquelles il prend un traitement par hydrochlorothiazide (Ésidrex®), amlodipione (Amlor®), et acétyl salicylate de lysine (Kardégic®). Il est très anxieux et prend régulièrement du bromazépam (Tranxène®).

'Mr. B, 71 year old, consults for abdominal pain and episodes of constipation. He also presents episodes of soft stools. His family doctor has talked to him about irritable bowel disease, which has not reassured him at all. He is a retired accountant. He has a history of arterial hypertension and arteritis for which he takes a treatment with hydrochlorothiazide (Ésidrex®), amlodipione (Amlor®), and lysine acetylsalicylate (Kardégic®). He is very anxious and takes bromazepam (Tranxène®) regularly.'

Table 14: Sample case of the Epreuves Classantes Nationales (ECN, 'National Classifying Examination'), and its translation; ECN were used for evaluating the vocabulary coverage of the system with regard to new clinical cases

normalisation and generation steps. We computed for each text the proportion of in-vocabulary and out-of-vocabulary words, and their average over all texts.

We also evaluated to what extent the ECN texts are different to the first cases used in system development. For this purpose, we compared the word types in the ECN cases to the word types in the initial cases and computed the percentage of ECN words that were not present in the development cases.

5.6.2 Results

Cases in our development set included 1,504 tokens (428 types), and two cases had several consultations. We evaluated the system's vocabulary coverage of the 169 new cases found in ECN texts (24,521 tokens, 4,112 types). These counts do not include dates and numbers, which we removed from these texts because they are not managed through dictionaries and hence raise no coverage problem.

We measured that 3,805 (92.5 per cent) of the word types occurring in the ECN texts did not occur in the development cases. This shows that the ECN texts used to test vocabulary coverage are really different from the cases used in system development.

The system's resources recognised an average of 97.85 per cent of ECN word types (stdev=1.91). A very low percentage of ECN word types were out of vocabulary (OOV) items (2.15 per cent), with an average of 2.34 per case (stdev=2.72).

Figure 11 shows that missing terms were mostly acronyms and abbreviations (e.g.

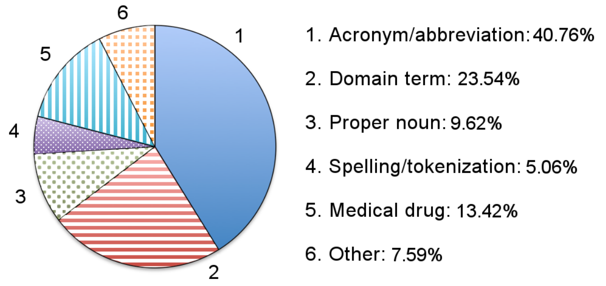


Fig. 11: Breakdown of types of out-of-vocabulary words in ECN texts

rcp), domain terms (e.g. *decubitus*) and medications (e.g. *zopiclone*). Other OOV words were names of people or locations (e.g. *France*), other words (e.g. *rendez-vous*), and spelling errors (**oédème* instead of *oedème*, 'edema'). Note that names of people and locations are not useful in the system, and are thus not included in our lexicons. Accordingly, the actual rate of OOV words is even lower than measured: resources covered an average of 98.09 per cent of ECN word types (stdev=1.86), and the average of OOVs per case in the ECNs was of 2.14 (stdev=2.67).

5.7 Qualitative analysis of vocabulary usage

5.7.1 Methods

To illustrate the difficulties in managing the variety of terms in our task, we provide a qualitative analysis of domain term usage in the interaction data. First, we analysed how medical evaluators used domain terms in dialogue logs. We only analysed dialogues in cases tested by more than one medical doctor (28 different cases). As an illustration, we focused on terms related to entity types of specific references to diseases (disease_spec) and symptoms (symptom_vb, which labels verb forms, and symptom). We examine here the actual occurrence of the terms observed in the interaction logs. Second, we analysed the term distribution in the corresponding VP records (only for those 28 cases). We did not include stop words nor the record section containing demographic data (e.g. proper names or civil status). For both aspects of the analysis, we obtained frequencies of usage of each item, and plotted them to examine their distribution across user interactions or records.

5.7.2 Results

In the interaction logs and patient records of the 28 analysed cases, we found 1,408 different tokens. Less than 60 (∼4 per cent) have a frequency over 10 and around 30 (∼2 per cent) occur in at least 10 cases. Figure 12 summarises our results. The frequency distribution of the terms for symptoms and diseases are represented in two plots above (A & B). Terms were analysed in the dialogues of all VP records (plot A) and we also counted the number of different cases where each term occurred in the interaction dialogues (plot B). Secondly, we plot the analysis of tokens in all

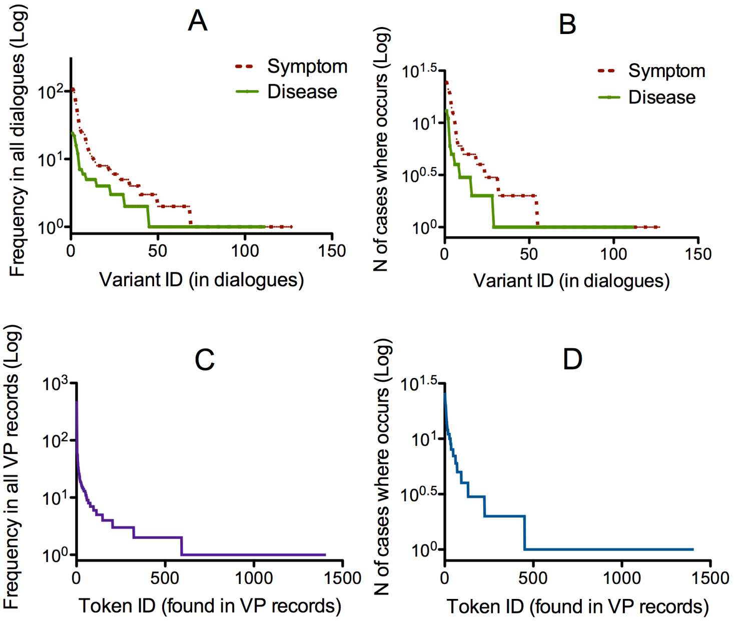


Fig. 12: Qualitative analysis of term distribution in dialogues and VP records

records: token frequency in all records (plot C), and the number of cases where each token appeared (plot D). All plots show that both token frequency and terms of these entity types follow a Zipfian-style distribution. Term management in our system needs to cope with a distribution of Large Number of Rare Events.

A quick look showed that very low-frequency items are domain terms (e.g. *cianotic* or *acetylleucine*). This illustrates the difficulty of term detection in our task and justifies our approach based on rich terminological resources. The methods we propose enhance the system's ability to adapt to different cases and detect rare and unseen vocabulary items for a successful interaction. Table 15 breaks down the most frequent terms for specific references to symptoms (labels `symptom` and `symptom_vb`) and diseases (`disease_spec`) observed in the dialogues; and Table 16 reports the frequency analysis of tokens in the records.

5.8 Discussion

Our system is aimed at dealing with new cases and therapeutic areas. To achieve that, we rely on robust and comprehensive terminology components, which, as far as we know, are unparalleled in current VP systems (Maicher *et al.* 2017). The method which consists in crowd-sourcing the evaluation of the answers of the dialogue system and the extending its question-answer database with the correct answer (Rossen, Lind and Lok 2009) does not seem scalable for extending the components of our system and its terminology. Because these variation phenomena

	Most frequent			Least frequent (selection)		
	Term	**Freq**	**# cases**	**Term**	**Freq**	**# cases**
Symptoms	*ache*	112	25	*cramps*	1	1
	pain	96	14	*cyanotic*	1	1
	pains	65	19	*itching*	1	1
	fever	45	21	*distress*	1	1
	lost weight	29	12	*difficulty urinating*	1	1
Diseases	*injury*	24	1	*abuse*	1	1
	hypertension	22	13	*accident*	1	1
	diabetes	16	11	*cerebrovascular accident*	1	1
	psoriasis	12	2	*accidents*	1	1
	pancreatitis	7	1	*alzheimer*	1	1

Table 15: Most and least frequent terms (translated from French) observed in dialogues with 28 different VPs for entity types of symptoms and diseases. Note that some terms cannot always be assigned to symptom or disease, and the dialogue context or VP record are needed to make a distinction. For example, *vertigo* is most commonly a symptom, but it might be a chronic condition in a specific case.

Most frequent tokens			Least frequent tokens (selection)		
Token	Freq	# cases	Token	Freq	# cases
patient (masc)	483	21	*abdomen*	1	1
patient (fem)	199	10	*accelerates*	1	1
years	120	26	*acetate*	1	1
months	60	19	*acetylleucine*	1	1
day	57	23	*acyclovir*	1	1

Table 16: Most and least frequent tokens (translated from French) observed in the 28 analysed VP records (only a selection of the least frequent tokens is shown)

are easier to process through thesauri, we rely on domain lexicons and ontological knowledge. Our approach is closer to that based on a taxonomy of questions (Talbot *et al.* 2016); however, we use terminologies available in the UMLS Metathesaurus.

Evaluation outcomes showed that the core lexical and terminological components seem stable and able to process new clinical cases. From a quantitative point of view, the NLU module achieved an F-measure of 95.8 per cent, balancing precision (96.8 per cent) and recall (94.9 per cent) when annotating entities in user input.

The test of vocabulary coverage brought out the few types of terms that occurred in unseen patient cases and were still missing in our lexicon: these only represented 2.16 per cent of terms in a collection of 169 descriptions of clinical cases, and most missing terms were acronyms and abbreviations. From a qualitative point of view, users who tested the system did not mention any error related to terminology needs.

According to the evaluations, the most important causes of failures in the dialogue manager are beyond terminology needs and might define the limits of a purely rule- and frame-based system. The difficulty of the task also accounts for processing failures of follow-up queries after general questions, especially in cases of missing information in the record. Medical doctors tend to ask general questions to begin to circumscribe a diagnosis; then, if the patient replies with the searched bit of information, they ask for more details. This requires both processing correctly the implicit information in the dialogue context and foreseeing all details to be queried: e.g. observations, descriptions (e.g. intensity) or temporal data related to a condition. The lack of pre-existing task-specific dialogue data hinders achieving a comprehensive coverage of question types, query variants and interaction contexts.

We would like to improve the naturalness of replies, especially those with long sentences and negative symptoms. The realism of responses depends on technical aspects as well as on how medical instructors input data.

The methods we propose should be valid for other dialogue tasks in other domains where rich lexical and/or ontological resources exist and a semi-structured database is available. That makes it possible to develop a system to collect interaction data, which can then be used in statistical or machine-learning approaches. Nonetheless, a key takeaway is the fact that, even when rich resources exist, these need an extensive effort of iterative filtering and task adaptation before production mode. As we explained, we created *ad-hoc* lexicons with lay variants or equivalences between noun terms and verbs (which are missing in the UMLS). Similar needs were reported when adapting the UMLS for concept indexing (Nadkarni, Chen and Brandt 2001).

In terms of dialogue management, we contribute with an approach for querying the database at each dialogue move by considering the semantic content of each dialogue state.

6 Conclusion

In this paper we highlighted the difficulties raised by the terminological needs of a dialogue system that aims at providing natural language interaction in a simulated medical consultation context, robust enough for multiple clinical cases. We designed three models involved in such a dialogue task: a patient record model, the knowledge model for the task, and a termino-ontological model.

This work focused on the termino-ontological model, which manages terminological and linguistic resources. To populate the model, we adopted a comprehensive approach to lexicon and terminology collection. We collected term variants for domain concepts based on existing medical terminologies, which helped us structuring terms according to the concepts they describe. We compiled large dictionaries of inflectional and derivational word variants. These resources enabled the system to: 1) recognise a large number of entities in the NLU step; 2) handle general and specific entities in the NLU or dialogue manager modules; 3) perform entity linking, entity normalisation and hierarchical reasoning; and 4) give priority to lay variants in the generation step. The quality of the collected resources allowed the system to obtain a high vocabulary coverage when tested on a large number of unseen cases: the system proved stable for the task and robust enough to cope with the vocabulary of new cases. Our system stands out from current research on VPs by its ability to handle a large variety of clinical specialities and cases. We developed the system with 35 different records from 18 medical specialities.

A total of 32 non-medical users and 39 medical students and doctors evaluated the system. Overall, the majority of users evaluated it as good or very good in most dimensions. The evaluation also highlighted aspects that deserve further work. User comments in the evaluation reflected these shortcomings, especially regarding follow-up utterances after general questions, improving the naturalness of some replies and handling missing information in the patient record.

We make the evaluation corpus used in this work available for the community.[13] In the context of a lack of dialogue resources, especially in the medical domain, we believe these data will be useful for moving ahead in the field.

Now that we have collected interaction corpora, we are focusing our research on machine-learning based methods. Specifically, we have begun exploring the classification of question types according to the system needs, i.e. the system's current rules or an alternative processing strategy (Campillos-Llanos, Rosset and Zweigenbaum 2017). Designing such a fallback strategy is our research interest, with a view to providing a satisfactory answer when a question cannot be handled with the current approach. We estimate that a subset of system's wrong and not-understood replies (overall, less than 20 per cent of replies) would need a fallback strategy.

The system has been adapted to English and Spanish, following the same design procedures and models (e.g. entity types scheme) as explained. Domain lists contain over 116,000 terms in English and 103,000 in Spanish; and dictionaries gather over 1,886,000 word/concept entries in English and 1,428,000 in Spanish. A thorough data collection and evaluation are needed to improve these versions of the system.[14]

7 Acknowledgements

This work was funded by BPI through the FUI Project PatientGenesys (F1310002-P) and by SATT Paris Saclay through project PVDial. We greatly thank all the evaluators who tested and assessed the system, and we thank VIDAL®, who provided a domain corpus of medications and disorders. We also thank Brigitte Grau and the anonymous reviewers for their valuable remarks on the manuscript.

[13] Available at: `https://pvdial.limsi.fr/data/PG-logs-eval.zip`

[14] These versions are available at: `www.audiosurf.net/pg_2018/select_case.php`

References

Bates, B., and Bickley, L. S. 2014. *Guide de l'examen clinique – Nouvelle édition 2014*. London/Montrouge: Arnette-John Libbey Eurotext.

Beveridge, M., and Fox J. 2006. Automatic generation of spoken dialogue from medical plans and ontologies. *Journal of biomedical informatics* 39(5): 482–499.

Benedict, N. 2010. Virtual patients and problem-based learning in advanced therapeutics. *American journal of pharmaceutical education* 74(8), article 143.

Bickmore, T. 2015. Conversational agents for automated inpatient and outpatient health counseling. In *Proc. of the AMIA Symposium*, San Francisco, USA, p. 2131.

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

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research* 32(suppl 1): D267–D270.

Bouamor, D., Campillos-Llanos, L., Ligozat, A.-L., Rosset, S., and Zweigenbaum, P. 2016. Transfer-based learning-to-rank assessment of medical term technicality. In N. Calzolari *et al.* (eds.), *Proc. of LREC 2016*, Portorož, Slovenia, pp. 2312–2316.

Campillos-Llanos, L., Bouamor, D., Bilinski, E., Ligozat, A.-L., Zweigenbaum, P., and Rosset, S. 2015. Description of the PatientGenesys dialogue system. In *Proc. of SIGDIAL*, Prague, Czech Republic, pp. 438–440.

Campillos-Llanos, L., Bouamor, D., Zweigenbaum, P., and Rosset, S. 2016. Managing linguistic and terminological variation in a medical dialogue system. In N. Calzolari *et al.* (eds.), *Proc. of LREC 2016*, Portorož, Slovenia, pp. 3167–3173.

Campillos-Llanos, L., Rosset, S., and Zweigenbaum, P. 2017. Automatic classification of doctor-patient questions for a virtual patient record query task. In *Proc. of the 16th BioNLP 2017 Workshop*, Vancouver, Canada, pp. 333–341.

Celikyilmaz, A., Deng, L., and Hakkani-Tur, D. 2017. Deep Learning for Spoken and Text Dialog Systems. In L. Deng and Y. Liu (eds) *Deep Learning in Natural Language Processing*, Berlin: Springer, pp. 49–78.

Cole, R. 1999. Tools for research and education in speech science. In *Proc. of the International Conference of Phonetic Sciences*, San Francisco, USA, vol. 1, pp. 277–281.

Cook, D. A., Erwin, P. J., and Triola, M. M. 2010. Computerized virtual patients in health professions education: a systematic review and meta-analysis. *Academic Medicine* 85(10): 1589–1602.

Coudé, C., Coudé, F.-X., and Kassmann, K. 2011. *Guide de conversation médicale - français-anglais-allemand*. Paris: Lavoisier.

Courtois, B. 1990. Un système de dictionnaires électroniques pour les mots simples du français. *Langue française* 87(1): 11–22.

Danforth, D. R., Procter, M., Chen, R., Johnson, M., and Heller, R. 2009. Development of virtual patient simulations for medical education. *Journal For Virtual Worlds Research* 2(2): 4–11.

Datta, D., Brashers, V., Owen, J., White, C., and Barnes, L. 2016. A Deep Learning Methodology for Semantic Utterance Classification in Virtual Human Dialogue Systems. In *Proc. of the International Conference on Intelligent Virtual Agents 2016*, Berlin: Springer-Verlag, pp. 451–455.

Dickerson, R., Johnsen, K., Raij, A., Lok, B., Hernandez, J., Stevens, A., and Lind, D. S. 2005. Evaluating a script-based approach for simulating patient-doctor interaction. In *Proc. of the Intern. Conference of Human-Computer Interface Advances for Modeling and Simulation*, pp. 79–84.

Donnelly, K. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics* 121: 279–90.

Ellaway, R., Candler, C., Greene, P., and Smothers, V. 2006. An architectural model for MedBiquitous virtual patients. `http://groups.medbiq.org/medbiq/display/VPWG/MedBiquitous+Virtual+Patient+Architecture`. Accessed 23 April 2018.

Epstein, O., Perkin, D., Cookson, J., and de Bono, D. P. 2015. *Guide pratique de l'examen clinique.* Paris: Elsevier Masson.

Galibert, O. 2009. *Approaches and methodologies for automatic Question-Answering in an open-domain, interactive setup.* Phd dissertation, Université Paris Sud - Paris XI.

Giorgino, T., Azzini, I., Rognoni, C., Quaglini, S., Stefanelli, M., Gretter, R., and Falavigna, D. 2005. Automated spoken dialogue system for hypertensive patient home management. *International Journal of Medical Informatics* 74(2): 159–167.

Gokcen, A., Jaffe, E., Erdmann, J., White, M., and Danforth, D. 2016. A corpus of word-aligned asked and anticipated questions in a virtual patient dialogue system. In N. Calzolari *et al.* (eds.), *Proc. of LREC 2016*, Portorož, Slovenia, pp. 3174–3179.

Hathout, N., Namer, F., and Dal, G. 2002. An experimental constructional database: the MorTAL project. In N. Hathout, F. Namer, and G. Dal (eds.) *Many morphologies*, Somerville, MA: Cascadilla Press, pp. 178–209.

Hoxha, J., and Weng, C. 2016. Leveraging dialog systems research to assist biomedical researchers' interrogation of Big Clinical Data. *Journal of biomedical informatics* 61: 176–184.

Hubal, R. C., Kizakevich, P. N., Guinn, C. I., Merino, K. D., and West, S. L. 2000. The virtual standardized patient. *Studies in health technology and informatics* 70: 133–138.

Hubal, R. C., Deterding, R. R., Frank, G. A., Schwetzke, H. F., and Kizakevich, P. N. 2003. Lessons learned in modeling virtual pediatric patients. *Studies in health technology and informatics* 94: 127–130.

Jin, L., White, M., Jaffe, E., Zimmerman, L., and Danforth, D. 2017. Combining CNNs and Pattern Matching for Question Interpretation in a Virtual Patient Dialogue System. In *Proc. of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, Copenhagen, Denmark, pp. 11–21.

Jokinen, K., and McTear, M. 2009. *Spoken dialogue systems.* Synthesis Lectures on Human Language Technologies, 2. San Rafael, CA: Morgan and Claypool Publishers.

Kenny, P., Parsons, T. D., Gratch, J., and Rizzo, A. A. 2008. Evaluation of Justina: a virtual patient with PTSD. In H. Prendinger, J. Lester, and M. Ishizuka (eds.), *Proc. of Intelligent Virtual Agents*, Berlin: Springer-Verlag, pp. 394–408.

Kenny, P., and Parsons, T. 2011. Embodied conversational virtual patients. In D. Perez-Marín and I. Pascual Nieto (eds.) *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*, Hershey: IGI Global, pp. 254–281.

Lelardeux, C., Panzoli, D., Alvarez, J., Galaup, M., and Lagarrigue, P. 2013. Serious game, simulateur, serious play : état de l'art pour la formation en santé. In *Actes du colloque Serious Games en Médecine et Santé (SeGaMED) 2013*, Nice: e-virtuoses, pp. L3/27–38.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10(8): 707–710.

Maicher, K., Danforth, D., Price, A., Zimmerman, L., Wilcox, B., Liston, B, Cronau, H., Belknap, L., Ledford, C., Way, D., Post, D., Macerollo, A., and Rizer, M. 2017. Developing a Conversational Virtual Standardized Patient to Enable Students to Practice History-Taking Skills. *Simulation in Healthcare* 12(2): 124–131.

Makhoul, J., Kubala, F., Schwartz, R., and Weischedel, R. 1999. Performance measures for information extraction. In *Proc. of DARPA Broadcast News Workshop*, Virginia, USA, pp. 249–252.

McCray, A. T., Srinivasan, S., and Browne, A. C. 1994. Lexical methods for managing variation in biomedical terminologies. In *Proc. of Annual Symposium Computer Applic. Medical Care*, Washington, pp. 235–239.

McCray, A. T., Burgun, A., and Bodenreider, O. 2001. Aggregating UMLS semantic types for reducing conceptual complexity. *Studies in health technology and informatics* 84: 216–220.

McTear, M., O'Neill, I., Hanna, P., and Liu, X. 2005. Handling errors and determining confirmation strategies—an object-based approach. *Speech Communication* 45(3): 249–269.

Nadkarni, P., Chen, R. and Brandt, C. 2001. UMLS concept indexing for production databases: a feasibility study. *Journal of the American Medical Informatics Association*, 8, 1, pp. 80–91.

Namer, F. and Zweigenbaum, P. 2004. Acquiring meaning for French medical terminology: contribution of morphosemantics. In *Proc. of the 11th MEDINFO Conference*, San Francisco, USA, pp. 535–539.

Nirenburg, S., Beale, S., McShane, M., Jarrell, B., and Fantry, G. 2008. Language understanding in Maryland virtual patient. In *Proc. of the 22nd International Conference on Computational Linguistics*, pp. 36–39.

Nirenburg, S., McShane, M., Beale, S., and Jarrell, B. 2008. Adaptivity in a multi-agent clinical simulation system. In *Proc. of AKRR'08, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, pp. 17–19.

Nirenburg, S., McShane, M., and Beale, S. 2009. A unified ontological-semantic substrate for physiological simulation and cognitive modeling. In *Proc. of International Conference on Biomedical Ontology (ICBO)*, Buffalo, New York, pp. 139–142.

Norvig, P. 2007. How to write a spelling corrector. `http://norvig.com/spell-correct.html`. Accessed 23 April 2018.

Paek, T. 2001. Empirical methods for evaluating dialog systems. In *Proc. of the workshop on Evaluation for Language and Dialogue Systems-Volume 9*, Toulouse, France, pp. 1–9.

Pastore, F. 2015. *How can I help you today? Guide de la consultation médicale et paramédicale en anglais.* Paris: Ellipses.

Patrick, J. and Li, M. 2012. An ontology for clinical questions about the contents of patient notes. *Journal of Biomedical Informatics* 45(2): 292–306.

Pinault, F. 2011. *Apprentissage par renforcement pour la généralisation des approches automatiques dans la conception des systemes de dialogue oral.* PhD dissertation, Avignon University, Avignon, France.

Purver, M., Ginzburg, J., and Healey, P. 2003. On the means for clarification in dialogue. In J. van Kuppevelt and R. W. Smith (eds.) *Current and new directions in discourse and dialogue*, Dordrecht: Springer, pp. 235–255.

Quirk, R., Crystal, D., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A comprehensive grammar of the English language.* New York: Longman.

Rombauts, N. 2014. *Patients virtuels: pédagogie, état de l'art et développement du simulateur Alphadiag.* PhD dissertation, Faculty of Medicine, Claude Bernard University, Lyon, France.

Rossen, B., Lind, S., and Lok, B. 2009. Human-centered distributed conversational modeling: Efficient modeling of robust virtual human conversations. In Z. Ruttkay *et al.* (eds.) *Proc. of the International Workshop on Intelligent Virtual Agents*, Berlin: Springer, pp. 474–481.

Rossen, B., and Lok, B. 2012. A crowdsourcing method to develop virtual human conversational agents. *International Journal of Human-Computer Studies* 70(4): 301–319.

Rosset, S., Galibert, O., Illouz, G., and Max, A. Integrating Spoken Dialog and Question Answering: the Ritel Project *Proc. of InterSpeech 2006*, Pittsburgh, USA, pp. 1914–1917.

Rosset, S., Galibert, O., Adda, G., and Bilinski, E. 2008. The LIMSI participation in the QAst track. In *Advances in Multilingual and Multimodal Information Retrieval*, Berlin: Springer-Verlag, pp. 414–423.

Roy, B., and Graham, T. N. 2008. *Methods for evaluating software architecture: A survey.* Technical Report 545, School of Computing, Queen's University at Kingston, Ontario, Canada.

Salazar, V. L., Eisman Cabeza, E. M., Castro Peña, J. L., and Zurita, J. M. 2012. A case based reasoning model for multilingual language generation in dialogues. *Expert Systems with Applications* 39(8): 7330–7337.

Siregard, P., Julen, N., and Lessard, Y. 2013. Apprendre le raisonnement clinique par jeu sérieux. In *Actes du colloque Serious Games en Médecine et Santé (SeGaMED) 2013*, Nice: e-virtuoses, pp. 79–83.

Stevens, A., Hernandez, J., Johnsen, K., Dickerson, R., Raij, A., Harrison, C., DiPietro, M., Allen, B., Ferdig, R., Foti, S., *et al.* 2006. The use of virtual patients to teach medical students history taking and communication skills. *The American Journal of Surgery* 191(6): 806–811.

Sutton, R. S., and Barto, A. G. 1998. *Reinforcement learning: An introduction.* Cambridge: MIT press.

Talbot, T. B., Sagae, K., John, B., and Rizzo, A. A. 2012a. Sorting out the virtual patient: how to exploit artificial intelligence, game technology and sound educational practices to create engaging role-playing simulations. *International Journal of Gaming and Computer-Mediated Simulations* vol. 4(3): 1–19.

Talbot, T. B., Sagae, K., John, B., Rizzo, A. A., and Playa, C. 2012b. Designing useful virtual standardized patient encounters. In *Proc. of the Interservice/Industry Training, Simulation and Education Conference*, 4(3), 3–6.

Talbot, T. B., Kalisch, N., Christoffersen, K., Lucas, G., and Forbell, E. 2016. Natural language understanding performance and use considerations in virtual medical encounters. *Studies in health technology and informatics* 220: 407–413.

Traum, D. R., and Larsson, S. 2003. The information state approach to dialogue management. In J. van Kuppevelt and R. W. Smith (eds.) *Current and new directions in discourse and dialogue*, Dordrecht: Springer, pp. 325–353.

Traum, D. R., Robinson, S., and Stefan, J. 2004. Evaluation of a multi-party virtual reality dialogue interaction. In *Proc. of LREC 2004*, Lisbon, Portugal, pp. 1699–1702.

van Schooten, B., Rosset, S., Galibert, O., Max, A., op den Akker, R., and Illouz, G. 2007. Handling speech input in the Ritel QA dialogue system. In *Proc. of Interspeech*, Antwerp, Belgium, pp. 126–129.

Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *Proc. of the 8th Conference of the European chapter of the Association for Computational Linguistics*, Madrid, Spain, pp. 271–280.

Young, S. J. 2006. Using POMDPs for dialog management. In *Proc. of Spoken Language Technology Workshop*, Palm Beach, Aruba, pp. 8–13.

Zweigenbaum, P., Baud, R. H., Burgun, A., Namer, F., Jarrousse, É., Grabar, N., Ruch, P., Le Duff, F., Forget, J.-F., Douyère, M., and Darmoni, S. 2005. A unified medical lexicon for French. *International Journal of Medical Informatics* 74(2–4): 119–124.